UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO DE FIN DE GRADO

CARACTERIZACIÓN DE LA HORA VALLE Y APLICABILIDAD EN LA RED ACADÉMICA ESPAÑOLA

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

Sergio Albandea Martínez

Junio 2015

CARACTERIZACIÓN DE LA HORA VALLE Y APLICABILIDAD EN LA RED ACADÉMICA ESPAÑOLA

AUTOR: Sergio Albandea Martínez
TUTOR: José Luis García Dorado
PONENTE: Javier Aracil Rico

High Performance Computing and Networking Research Group (HPCN)
Dpto. Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio 2015

Resumen

Las tareas de tipo *bulk*, actividades de gran coste computacional o transferencias entre centros de datos que implican gran cantidad de datos pero con bajos requisitos de latencia, tienen una incidencia directa en la operación de una red, así como una relevancia muy importante en el coste. Ejemplos de tareas de tipo *bulk* son distribución de bases de datos, replicación de recursos o *backups* de seguridad.

Cómo planificar tales transferencias a lo largo de un día es un problema que ha recibido mucha atención. Destacan las propuestas que, recientemente, intentan sacarle partido a las peculiaridades del modelo de pago por ancho de banda más habitual, el modelo percentil 95. Transmitiendo por debajo de ese umbral se consiguen descuentos significativos.

En este trabajo, definimos el estudio de la hora valle como solución para este problema, pues efectuando dichas transferencias en las horas de menor utilización de una red o subred se consigue disminuir su impacto. Estas ideas han sido puestas en práctica en medidas reales de la red académica española, RedIRIS.

Primero, hemos examinado y verificado la integridad de las medidas del sistema de monitorización basado en flujos desplegado en RedIRIS. El estudio ha revelado peculiaridades en las subredes, tanto por fallos en las sondas de toma de medidas como en su propia transmisión o su recolección. Esto ha llevado a seleccionar catorce subredes, analizadas en un periodo de tiempo amplio -seis años-, lo que se traduce en un marcado carácter generalista.

A continuación, hemos modelado la hora valle como una variable gaussiana, con objeto de poder comparar el comportamiento de los distintos exportadores, así como predecirla mediante factores externos como el día de la semana, el mes o la condición de día laboral o festivo.

De este modo, podemos encontrar un instante óptimo para la realización de este tipo de tareas *bulk*, tanto en el marco de una única subred como en el caso de una ruta que atraviesa varias de ellas. En concreto para esto último, presentamos distintas alternativas, como el uso de almacenamiento interno en cada salto, un promedio de las distribuciones individuales o la agregación de varios de estos puntos en *metanodos*.

Palabras clave

Hora valle, transferencias bulk, percentil 95, centros de datos, planificación de red, medidas de red, flujos de red, RedIRIS, ancho de banda, ANOVA.

Abstract

Bulk transfers, defined as activities with high computational cost or transfers between data centers that involve large amounts of data but with low latency requirements, have a direct effect on the performance of a network, as well as a very important significance on its cost. Examples of such bulk tasks are database distribution, resources replication and security backups.

How to plan such transfers along the day is an issue that has received much attention, We remark those proposals that recently have proposed to exploit the peculiarities of the 95-th percentile charging model –the most popular in Internet. Such proposals explain how transmitting precisely below the 95 threshold gives significant cost savings.

In this work, we define the valley-hour study as a solution to this problem. This is because making such transfers during hours of lower use of a network allows reducing their impact. These ideas have been put into practice in the Spanish Research and Education Network, RedIRIS.

First of all, we have examined and verified the integrity of the measurements according to the monitoring system based on NetFlows built on RedIRIS. The study has revealed peculiarities in several points of presences (PoP). Essentially, problems with the probes, the transmission and also with the collection of the measurements. This has led to select fourteen exporters, analyzed during a six-year period, resulting in a strong generalist nature.

Then, we have modeled the valley-hour as a Gaussian variable in order to compare the behavior of this feature in different PoPs, and predict it by external factors such as weekday, month or condition of day distinguishing between holiday or working day.

Therefore, we can find an optimal moment to perform such bulk tasks, both within a single PoP and a path that cross multiple PoPs. In particular for the latter, we present some alternatives, including the use of internal storage at each PoP, an average of individual distributions or the aggregation of several of these points in meta-nodes.

Keywords

Valley-hour, bulk transfers, 95-percentile method, data centers, network planning, network measurement, NetFlows, RedIRIS, network bandwidth, ANOVA.

Agradecimientos

Con este texto y todo el trabajo desarrollado tras estas líneas acaba una etapa, en la que, como en el resto de aspectos de la vida, el verdadero disfrute se encuentra en el camino que te lleva al objetivo… y en las personas que de una manera u otra forman parte de él.

Agradezco a mi tutor en este trabajo, José Luis, su confianza y haberme dado la oportunidad de realizarlo dando facilidades desde el primer hasta el último día. Seguramente, con el tiempo guardaré como un bonito recuerdo esas largas tardes e incluso noches en el despacho, rebatiendo sobre distribuciones, modelos y demás asuntos con solo un bolígrafo y varias hojas de papel. He aprendido mucho.

Gracias a Jorge -quién me iba a decir que acabaría mencionando en este texto a ese desconocido que sentó a mi lado aquel lejano primer día-, a Guille y a Álvaro por haber compartido todos estos años. Prácticas interminables, apuntes, estudio en la biblioteca y en la *ofi*, pero también muchísimas idas de olla y algunos grandes momentos fuera de la EPS. Me lo he pasado genial con vosotros. No me olvido de Deinis, sobre todo por estos últimos meses, y del resto de compañeros con los que he tratado en alguna ocasión.

Igualmente, quiero acordarme de todos aquellos que, lejos de la universidad, han contribuido no solo a que haya conseguido concluir este grado, sino que también a que haya podido vivir grandes y enriquecedores momentos fuera de la carrera.

Un enorme gracias a mis padres por poner los medios para que estudiara, por no presionarme y haber confiado plenamente en que, algún día, lograría alcanzar este último paso. También a mi hermana.

Y por supuesto que no, no me he olvidado de ti, Cristina. Estaba claro que teníamos que acabar a la vez. Gracias por todo lo que hemos vivido y por lo que, espero, esté por venir.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

Los volúmenes de datos transferidos usando la infraestructura de Internet han crecido día a día desde sus inicios. Si en un principio se encontró que una furgoneta cargada de discos duros podía suponer una competencia real para el intercambio de información, la creciente popularidad de Internet, su dispersión y estandarización, así como la bajada de precios, han hecho de las transferencias de datos uno de los principales pilares del Internet que hoy en día conocemos. Y no solo por las transferencias que realizan los usuarios finales de Internet, sino también por las propias transferencias que las actividades de gestión y administración de las redes de comunicaciones o las aplicaciones sobre Internet involucran.

Esto es, la mayoría de las redes desplegadas en la actualidad requieren tareas diarias de consolidación, sincronización, compactación o distribución de bases de datos, replicación de recursos como máquinas virtuales o *backups* de seguridad, entre otras muchas tareas. Estas actividades -denominadas de tipo *bulk*- tienen en común dos características: por un lado, utilizan un gran ancho de banda al mover grandes volúmenes de datos; por otro, son tareas que normalmente pueden ser llevadas a cabo en cualquier instante a lo largo del día y no exigen un *timing* concreto [1].

Con intención de evaluar qué volúmenes implican estas tareas, los autores de [2] estudiaron el tráfico de decenas de centros de datos. Mostraron que un 77% de estos centros ejecutan *backups* y replicaciones de aplicaciones de forma diaria, entre más de tres localizaciones. Más de la mitad de los gestores de estos centros informaron de que tenían almacenados más de un PB de datos en sus localizaciones primarias. Y de manera significativa para este trabajo, el 70% de estos gestores estimaron entre 1 y 10 Gb/s la tasa media entre sus diversos puntos de presencia y la mitad de ellos, más de 5 Gb/s. Estas cifras se traducen en volúmenes entre 330 TB y 3'3 PB cada mes.

En este escenario, surge la necesidad de planificar tales transferencias de tipo *bulk*, de forma que minimicen el coste o, muy habitualmente, la operación habitual de la red. En este sentido, la comunidad académica que ha prestado atención con asiduidad a la caracterización y monitorización del tráfico en Internet [3] se ha centrado en los últimos años en el estudio de los momentos de mayor utilización [4]. Sin embargo, las transferencias de tipo *bulk* hacen necesario un análisis complementario centrado en los puntos de menor ocupación de las redes, pues éstos pueden ser aprovechados para realizar dichas transferencias.

Así, determinando los momentos valle -en este trabajo y de forma habitual, de duración una hora [5]-, se minimiza la interferencia con la operación habitual de la red y se aprovechan recursos ya desplegados e infrautilizados de manera puntual. Igualmente, ante un escenario de infraestructuras alquiladas, con costes derivados de la contratación de un proveedor de servicios de Internet (ISP), cuyo cobro por prestación se establece con frecuencia mediante un modelo de percentil 95 del uso de la red [6], también disminuye. Además, se consigue repartir el ancho de banda de forma más eficiente, lo que se traduce en una mayor calidad de servicio.

1.2 Objetivos

El objetivo de este trabajo es el estudio de la hora menos cargada como mecanismo para dar respuesta al momento ideal, o *timing*, del problema de transferencias *bulk*. Para ello, se propone el modelado y caracterización de la hora menos cargada y su generalización en una red de miles de usuarios repartidos en múltiples puntos de presencia a lo largo de un periodo de tiempo representativo.

En concreto, se estudiará la red académica española, RedIRIS, que da servicio a más de un millón de usuarios repartidos por toda España, y a la que se ha tenido acceso durante un periodo de seis años, desde 2008 hasta finales de 2013. Para ello, se efectuará un estudio tanto longitudinal como espacial de los datos extraídos mediante flujos de red que permita comparar las similitudes y diferencias entre puntos de presencia (PoPs) –en esencia, los distintos *routers* que componen la red- de RedIRIS.

De esta manera, se podrán plantear y conocer las peculiaridades de cada subred y determinar si se puede enunciar algún tipo de invariante para este problema, característica de gran utilidad para conformar un modelo matemático pero un reto en el heterogéneo Internet actual [3].

Para acometer este objetivo se establecen varias fases:

- **Estudio del estado del arte**. Se explicará el funcionamiento de la tecnología NetFlow que da lugar a las características del uso de red a través de los procesos que intervienen en el mismo. Se revisarán los esfuerzos en la comunidad de Internet por dar solución al problema del *timing* para las transferencias de tipo *bulk*, y qué propuestas se han hecho en este sentido. Se presentará nuestra propuesta y se pondrá en contexto con aproximaciones complementarias.

- **Visión general de los datos disponibles**. Se explicará qué datos hemos podido estudiar y se detallarán los pasos seguidos para analizar de manera preliminar cada subred o punto de presencia de RedIRIS. Mediante la visualización de las redes y la formulación de test empíricos, se debatirá la estacionariedad o no de los datos y, posteriormente, se indicarán los PoPs repartidos por la geografía española que van a formar parte del análisis.

- **Modelado**. Se comprobará la normalidad de la distribución de probabilidad de la hora valle. Se estudiará si es necesaria, y en ese caso, si es posible, la introducción de algún tipo de transformación que permita considerar una distribución gaussiana. En caso contrario, se optará por un modelo alternativo.

- **Predicción de la hora menos cargada**. Finalmente, se expondrá una propuesta para la predicción de la hora menos cargada y se examinará si dicha proposición es independiente para cada una de las subredes o si se puede extender a varios puntos de presencia.

1.3 Organización de la memoria

La memoria consta de los siguientes secciones o capítulos:

- **Sección 1: Introducción**. Motivación; objetivos; organización de la memoria.

- **Sección 2: Estado del arte**. Repaso a las técnicas que hacen posible este trabajo: características y funcionamiento de NetFlow; profundización en la motivaciones: problemática de *data centers* y transferencias *bulk*; justificación del estudio de la hora menos cargada como solución.

- **Sección 3: Datos disponibles**. Presentación de RedIRIS y su topología; particularidades de la monitorización basada en flujos y los ficheros en los que resulta; disponibilidad de las medidas a analizar.

- **Sección 4: Análisis preliminar**. Visualización del fenómeno; circunstancias que amenazan la integridad de los datos y tolerancia a errores; exportadores bajo estudio y sus perturbaciones; visualización de la estacionariedad de la variable hora valle; mirada a los factores externos que pueden condicionar el problema.

- **Sección 5: Caracterización**. Aproximaciones propuestas: distribución normal y distribución uniforme; estudio de ambas alternativas y comparativa; decisión y efecto.

- **Sección 6: Predicción**. Consideraciones; análisis ANOVA; predicción para una única subred; diversas formas de predicción para una ruta que pasa por varios exportadores.

- **Sección 7: Conclusiones y trabajo futuro**. Conclusiones tras la realización del estudio; planteamiento de posibles futuras publicaciones relacionadas.

2 Estado del arte

El estudio del estado del arte del momento hora valle se divide en tres partes. Primero, se describen las técnicas y herramientas desarrolladas con anterioridad por la comunidad de Internet y en las que se apoya nuestro estudio de la hora valle. De este modo, la primera parte ahonda en los procedimientos y tecnologías que posibilitan la extracción de los datos que serán examinados a lo largo de este trabajo.

En segundo lugar, se motiva este trabajo con artículos que han identificado la importancia que tienen las transferencias de tipo *bulk* en Internet y han propuesto distintas técnicas para planificar de manera inteligente cuándo y cómo realizar estas transferencias.

En tercer lugar, se describen distintos trabajos que han indagado en la importancia de caracterizar Internet. En concreto, destacamos la importancia de los invariantes en Internet; esto es, comportamientos que presenta una red y que son extrapolables a otras. Así, estudiaremos otros trabajos que, en su afán de identificar invariantes en Internet, han propuesto y usado distintos modelos para caracterizar y comparar comportamientos en la red.

En resumen, un primer paso consiste en explicar las tecnologías que han permitido obtener las medidas -en concreto, será la tecnología NetFlow-, motivar el problema que tratamos -*timing* de las transferencias *bulk*- y cómo ha sido tratado éste por la comunidad científica, con especial atención a la proposición NetStitcher. Finalmente, se detallan qué otros trabajos o estudios nos han servido para proponer la hora menos cargada -más concretamente, la importancia de la hora más cargada para el dimensionado de la capacidad de enlaces- como una buena aproximación al problema presentado.

2.1 NetFlow

2.1.1 Introducción

La tecnología usada en este trabajo para obtener medidas de red ha sido NetFlow [7]. NetFlow es un conjunto de herramientas, mecanismos y protocolos desarrollado por Cisco Systems que permite la creación y exportación de registros de flujos de red.

Se conoce como un flujo de red al conjunto de paquetes consecutivos que comparten las mismas IPs origen y destino, puertos origen y destino y protocolo, atributos que se denominan *quíntupla*. De este modo, NetFlow inspecciona el tráfico de una red, analiza los paquetes que por ella transcurre, y crea una tabla con información -esto es, un registro- de flujos. Tradicionalmente, la tabla contiene hasta siete campos: dirección IP origen, dirección IP destino, puerto origen, puerto destino, protocolo, clase de servicio (TOS) e interfaz lógica de entrada y salida [7].

La razón de incluir las interfaces radica en que NetFlow nació como un mecanismo auxiliar en el enrutado, de modo que únicamente se buscara en la tabla de reenvío el primer paquete del flujo y el resto le siguiera, pero pronto se cayó en la cuenta de sus múltiples otras utilidades.

La información de un flujo contiene datos de gran utilidad para extraer estadísticas, pues las direcciones IP origen y destino ofrecen la información de dónde se ha originado el tráfico y hacia dónde se dirige, los puertos pueden servir para deducir el tipo de aplicación de la que procede ese flujo y el tamaño medio del paquete da indicios de la carga de enrutado, por citar varios ejemplos [7].

En el pasado, NetFlow ha sido usado con éxito para monitorizar redes [8], tareas de gestión y seguridad [9] o clasificación de tráfico [10], entre muchas otras finalidades. De hecho, NetFlow se ha estandarizado, dando lugar al término IPFIX [11], que se refiere a cualquier flujo, sea o no de Cisco, y con una definición más flexible y completa del registro de red.

Frente a sistemas de monitorización alternativos basados en la captura de paquetes o agregados, que requieren grandes demandas de espacio en disco o altas capacidades de procesamiento, el interés en emplear NetFlow reside en que combina exigencias bajas tanto de almacenamiento como rendimiento [12]. Reducir ambos requerimientos es posible mediante el muestreo de los paquetes de entrada, que habitualmente se realiza de forma determinista, es decir, tomando para la extracción de sus características solamente uno de cada 'X' paquetes, siendo 'X' un número prefijado.

2.1.2 Funcionamiento

Un sistema de monitorización basado en NetFlow tiene tres elementos básicos. El primero de ellos es un grupo de *routers* con capacidad NetFlow, los cuales acceden al tráfico y van formando los registros NetFlow en una tabla *hash* implementada sobre memoria RAM de alta velocidad [13].

Esta tabla se denomina NetFlow Caché o simplemente tabla de flujos activos. En esta tabla, cada flujo está representado por un registro que es actualizado cada vez que se conmutan los paquetes pertenecientes al flujo, contabilizando los paquetes y los bytes por flujo o cualquier otra información que contenga el flujo.

La Figura 2-1 ilustra el proceso de formación y registro de los flujos de red.

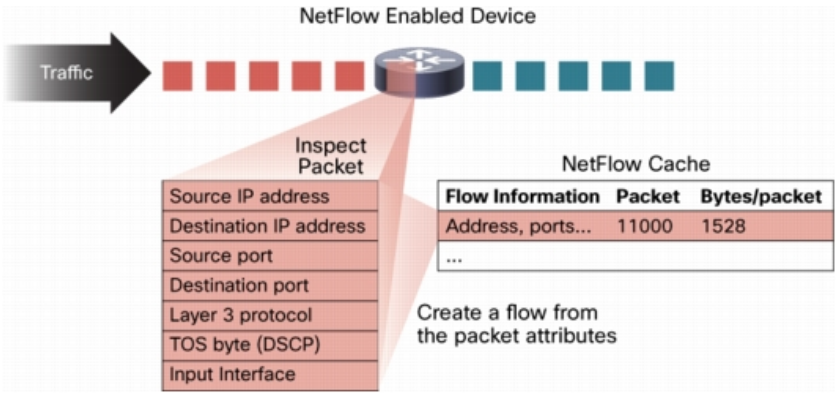

Figura 2-1: Funcionamiento de un *router* con capacidad NetFlow [7]

Aunque el tamaño que ocupan los registros de los flujos sea pequeño, estos se han de eliminar cada cierto tiempo, tras el cual se considerará que dicho flujo ha concluido. La

decisión de borrado puede producirse por distintos motivos [14]: si una conexión TCP finaliza o se resetea; si los flujos han estado inactivos por un tiempo fijo, normalmente 15 segundos; si la caché se llena; si el *router* se queda sin recursos o, en último caso, si transcurre un tiempo máximo determinado, normalmente 30 minutos. Cuando un flujo se considera que ha acabado, este se exporta y se elimina. Exportar un flujo consiste en enviar mediante el protocolo NetFlow que define esta tecnología -nótese que comparten nombre- el registro a un elemento de red.

De esta manera, definimos como segundo elemento de una arquitectura de monitorización basada en flujos de red un colector de NetFlows [15], el dispositivo que recibe los flujos y los almacena. El software más usado de libre acceso para almacenar de forma ordenada los registros es el conjunto de herramientas llamadas Flow-Tools [16].

En concreto, este software efectúa la escucha un puerto donde el *router* NetFlow debe enviarle los registros; Flow-Tools [16] comprime estos registros y los guarda en ficheros comprimidos en ciclos de 15 minutos. Estos registros son luego accesibles mediante la API que define Flow-Tools [16], con funciones para concatenar ficheros, filtrar por algunas características como por ejemplo IPs, incluidas las del propio exportador.

Finalmente, en la arquitectura es típico contar con un elemento analizador [17]. Este equipo o conjunto de ellos, accede a los flujos de red preparándolos de manera automática para su posterior análisis. Ejemplos de esto son medidas del ancho de banda, proporciones por IPs o por puertos, momentos de máxima o mínima carga o identificación de aplicaciones, aunque en este caso es razonable guardar también los primeros bytes del flujo como si de un tipo más de información se tratara.

2.2 Transferencias bulk

2.2.1 Concepto

Se denomina trasferencias *bulk* a aquellas transferencias entre equipos que tienen requisitos poco exigentes en cuanto a la latencia, pero que por el contrario implican grandes cantidades de datos [18]. Ejemplos de este tipo de tráfico son tareas diarias de consolidación, sincronización, compactación o distribución de bases de datos o replicación de recursos como máquinas virtuales o *backups* de seguridad, entre otras muchas.

Las transferencias de tipo *bulk* son muy habituales en *data centers* o centros de datos. Según algunas publicaciones como [2], en la que se estudió el tráfico de decenas de ellos, un alto porcentaje -77%- de estos centros efectúan *backups* y replicaciones de aplicaciones cada día, llevando a cabo conexiones con distintas localizaciones.

Además, estas tareas implican grandes usos de ancho de banda y fracciones del tráfico total de una red o subred. Este aspecto quedó patente en el mismo estudio [2], en el que se halló que más del 50% de los data centers contaban con más de un PB de almacenamiento. Para el 70% de ellos, la tasa media de utilización se situó entre 1 y 10 Gb/s, y, para el 50%, en cifras superiores a 5 Gb/s, lo cual se traduce en volúmenes considerables: por cada mes, entre 330 TB y 3'3 PB.

2.2.2 Planificación de las tareas de tipo bulk

En este escenario, han sido varias las ideas propuestas por la comunidad científica y tecnológica para intentar planificar las trasferencias de tipo *bulk* de la forma más provechosa posible, pensando tanto en el ahorro económico como en el reparto eficiente de la utilización del ancho de banda.

Los autores de [18] observaron que el coste del ancho de banda no es constante a lo largo del día. La razón radica en que el uso de la red es desigual, depende del momento del día en el que nos encontremos. Por ello, tanto en (i), una red con enlaces de coste fijo -muchas veces propietaria de la infraestructura- o tarifa plana, como en (ii), una red que paga por el uso de los enlaces, ciertas horas del día son más adecuadas para las transferencias de tipo *bulk*. En el primer caso, en tanto que hay momentos en los que el ancho de banda desplegado está simplemente infrautilizado; en el segundo, debido a que el pago del uso de los enlaces suele seguir modelos basados en utilización de picos.

En este último caso destaca que el mecanismo habitual por el que las compañías pagan por ancho de banda es aplicando el arquetipo denominado percentil 95. Por el modelo percentil 95, las compañías arrendadoras miden la demanda de ancho de banda, calculan el percentil 95 -típicamente a grano SNMP, esto es, 5 minutos- y cobran multiplicando el tiempo de utilización del enlace por este medida percentil 95. Así, las compañías que arriendan pagan en muchas ocasiones por anchos de banda que no utilizan de forma efectiva.

Es lógico razonar que un patrón de tráfico constante puede tener el mismo coste que otro irregular, siempre y cuando la muestra 95 sea equivalente, aun cuando el agregado del volumen transmitido sea radicalmente distinto en ambas situaciones. De ello deriva el interés del análisis que se desarrolla en este trabajo.

El argumento detrás de este tipo de modelo de pago se discute en [19]. Se basa en que el coste de desplegar infraestructura es proporcional a los picos de uso, pues estos determinan la capacidad del enlace, sin un impacto real del uso de la red en los momentos alejados de los picos. Sin embargo, los arrendadores de ancho de banda se preocupan de agregar perfiles entre usuarios dispares para sacar el máximo beneficio de la multiplexación estadística.

En dicho estudio propusieron la herramienta NetStitcher, intentando sacar partido a dos tipos de situaciones, permitiendo dos modos de funcionamiento dependiendo el tipo de red: en aquellas de tipo propietario o tarifa plana, plantean trasmitir datos de tipo *bulk* solo en ventanas de 3 horas de duración de madrugada; en aquellas redes que siguen modelos percentil 95, definen un mecanismo por el cual solo se transmiten en periodos por debajo de la medida 95, por tanto, sin modificarla y sin coste añadido.

No obstante, es posible que el volumen a transmitir sea mayor que el que permiten estas ventanas de tiempo, así como superior al límite que el percentil 95 impone. Consecuentemente, cada módulo NetStitcher no solo planifica, sino que tiene capacidades de almacenamiento. Dado el caso de no poder transmitir todos los datos, estos se retienen esperando una oportunidad de trasmitir.

Los autores de la publicación evaluaron los beneficios de este modelo, que se valoran como muy prometedores. Especialmente, destacaron los beneficios de transmitir del este

al oeste y de mover los datos siguiendo los husos horarios, siendo menos exitosas las transmisiones hacia el oeste al viajar hacia horarios más tardíos.

Asimismo, otros trabajos han seguido las pautas de NetStitcher y han sugerido mejoras en las que las transmisiones son divididas en bloques o *chuncks*, facilitando su transmisión por distintas rutas [20] o mostrando su utilidad en más redes [21].

2.3 Invariantes en Internet y su caracterización

Internet es, por definición, un conjunto de redes administradas por distintas instituciones que las gestionan y organizan de manera diversa. En adición a esto, el despliegue masivo de Internet y su llegada a cada rincón del mundo ha supuesto que los usuarios sigan patrones de uso distintos, con una gran heterogeneidad tanto de aplicaciones como de contenidos y también en el modo en el que se accede a ellos. Como añadido, este proceso está lejos de estabilizarse y cada día se añaden nuevos elementos al problema.

La respuesta a esta diversidad por la comunidad de Internet es la búsqueda de invariantes [3]. Un invariante referido a Internet es una característica del mismo que comparten varias de sus redes. Su identificación y, sobre todo, su extrapolación a la totalidad de Internet, es un auténtico reto debido a la gran escala y a la dificultad de tener medidas diversas en el tiempo y en el espacio [22].

Pese a las dificultades comentadas, conocer más a fondo Internet, hasta el punto de poder caracterizarlo, se ha convertido en una actividad fundamental para su optimización, predicción y dimensionamiento futuro. En consecuencia, han sido muchos los trabajos que han prestado atención a su estudio, analizando distintas características y constituyendo su modelado de la forma más general posible.

Algunos ejemplos son [4] [23] [24] [25], en los que, mediante la caracterización del ancho de banda, se pretende estimar las demandas futuras. En concreto, en [4] se estudia la relación de la hora más cargada -métrica de tipo pico de uso muy popular para dimensionar los enlaces y heredera de mecanismos del sistema telefónico tradicional- con la población de una red. La hora cargada es tratada como un sistema lineal entre la población y el tamaño de la red, permitiendo dar una estimación del uso de la red mediante variables independientes fácilmente medibles.

En concreto, se usa ANOVA, una técnica estadística que aplicaremos también en este trabajo y que consiste en, una vez cumplidos ciertos supuestos -como la normalidad de la distribución de la variable bajo estudio o la igualdad de las varianzas de las muestras que conforman el experimento-, una prueba o test de la significación o no de cada uno de los factores. El análisis mostró que la capacidad de los enlaces mostraban poca influencia en la hora cargada en las redes, mientras que la población sí se podía tomar como una variable explicativa.

Por su parte, en [23] sugieren un método mediante *wavelets* y series temporales lineales para predecir cuándo y dónde se deben llevar a cabo mejoras en una red IP. Los autores muestran que el tráfico de la red exhibe tendencias visibles a largo plazo, periodicidades fuertes y variabilidad en múltiples escalas de tiempo. El intento de determinación de invariantes mediante un análisis longitudinal se repite de forma similar en [24]. Este

estudio, que pone atención en las capas TCP e IP y en el uso de aplicaciones sobre MAWI *dataset*, una red sujeta a cambios en el ancho de banda y diversas anomalías, concluye una importante dependencia con el tiempo del número de bytes y del número de paquetes.

Por último, en [25] se emplea, al igual que en este estudio, flujos de red, con el objetivo de enunciar un modelo mixto para calibrar el efecto de las distribuciones *heavy-tailed*; en esencia, distribuciones con mayor probabilidad de la habitual en las zonas de baja probabilidad. Empleando ANOVA describen que la distribución geográfica es fuertemente dependiente de la red IP origen y subrayan el enorme nivel de agregación que se requiere para observar un patrón geográfico estable.

2.4 Contribuciones de este trabajo

En esta sección hemos profundizado acerca de NetFlow, el mecanismo que nos permite la extracción de los flujos de red y, en consecuencia, la obtención de los datos que sirven como base del análisis. También hemos percibido la problemática que suponen las transferencias *bulk* en cuanto a su repercusión en el cobro por el uso de la red cuando este depende del valor máximo de utilización.

Hemos sondeado varias publicaciones relacionadas con nuestro estudio y hemos conocido sus aproximaciones, que nos han guiado a pensar y a argumentar la introducción de la hora valle como propuesta para la optimización de estas transferencias. Su predicción, producto de los flujos de RedIRIS, tendrá en cuenta las bajas imposiciones en cuanto a *timing* que se establecen.

Posteriormente, hemos recalcado las dificultades que se ha encontrado la comunidad académica para declarar invariantes de Internet, motivadas por su continua expansión y los cambios y diferencias en cuanto al comportamiento de los usuarios. El hallazgo de algún tipo de característica intrínseca a varias subredes será igualmente uno de los propósitos marcados en este trabajo.

Fruto del repaso de las publicaciones que tratan invariantes de Internet, sospechamos que nuestro análisis probablemente depare heterogeneidad entre unas subredes y otras, y nos encontremos con escollos para establecer grupos con comportamientos semejantes. Para aseverar esta idea aplicaremos, entre otras herramientas, un análisis ANOVA a los datos, el cual nos ha resultado útil para presentar un modelo que determine cuándo deben realizarse las tareas de tipo *bulk*.

3 Datos disponibles

El desarrollo de este proyecto comienza introduciendo la red académica española RedIRIS, la que da lugar y sobre la que se basan los análisis posteriores. En concreto, se explicará su topología, compuesta de diecinueve puntos de presencia, esencialmente uno por comunidad autónoma. Estos puntos de presencia en ojos del sistema de monitorización consistirían en un *router* -aunque físicamente sean varios- con capacidad NetFlow. Por esto último, y como la información de monitorización que facilitan son los flujos exportados, serán frecuentemente denominados como *exportadores*.

A continuación se explicará el sistema de monitorización basado en NetFlow que está desplegado en RedIRIS bajo el proyecto DIOR / ANFORA [26]. De este modo, introduciremos el formato de los archivos binarios, comprimidos y de texto que genera el sistema. Entre las medidas extraídas estarán las que se utilizarán en este trabajo.

Finalmente, se mostrará, mediante una tabla, una panorámica de los datos disponibles, pues la evolución de RedIRIS-10 a RedIRIS-NOVA propició que varios de los puntos de presencia abandonaran progresivamente el proyecto y que, por tanto, dejaran de servir flujos de red para ser analizados.

3.1 RedIRIS

3.1.1 Introducción

El análisis y las conclusiones derivadas de este trabajo se basan en los datos disponibles en RedIRIS. RedIRIS es la red académica y de investigación española y proporciona servicios avanzados de comunicaciones a la comunidad científica y universitaria nacional, financiada por el Ministerio de Economía y Competitividad y gestionada por la entidad pública Red.es, dependiente del Ministerio de Industria, Energía y Turismo. En la actualidad, RedIRIS da servicio en más de 500 instituciones, principalmente universidades y centros públicos de investigación [27], lo que se traduce en más de un millón de usuarios.

El proyecto de RedIRIS se inició en 1988 -entonces bajo el nombre de IRIS, Interconexión de los Recursos Informáticos- con el objetivo de la implantación de una red que permitiera el intercambio de información y la coordinación entre grupos, esencialmente universidades, a semejanza de la comunidad académica en muchos países europeos [28].

3.1.2 Topología

Aunque hoy en día RedIRIS cuenta con una red de fibra denominada RedIRIS-NOVA que fue puesta en marcha progresivamente desde 2011, este trabajo estudia los flujos proporcionados por la red RedIRIS-10, en producción entre 2006 y 2012. RedIRIS-10 es una red IP dimensionada para soportar tráfico de propósito general y otros servicios que solventó los problemas de congestión que se daban en versiones anteriores [29].

RedIRIS-10 está dotada de una estructura troncal híbrida con enlaces a 10 Gb/s y mallada. Se conecta con el resto de puntos de Internet *comerciales*, así como con el GÉANT, *European Research and Education Network* [25]. RedIRIS-10 cuenta con diecinueve puntos de presencia distribuidos en las distintas comunidades autónomas, con el caso especial de Canarias, con dos PoPs, uno en Las Palmas y otro en Tenerife.

Estos puntos de presencia son, a alto nivel, un centro de datos que se encarga, entre otras funciones, de dar conectividad a todas las instituciones sitas en la comunidad autónoma -fundamentalmente universidades y hospitales- y conectarlas a Internet a través del punto neutro Expanix -organización sin ánimo de lucro que gestiona y mantiene un punto neutro de telecomunicaciones a nivel nacional formado por enlaces a 10 Gb/s y dos salidas internacionales a la misma velocidad [29]-, localizado en Madrid. Cataluña adicionalmente utiliza además el punto neutro de Catnix. La topología de RedIRIS-10 se muestra en la Figura 3-1.

Todos los *routers* de cada punto de presencia tienen capacidades NetFlow; su explotación con fines de monitorización son explicados en la siguiente Subsección 3.2.

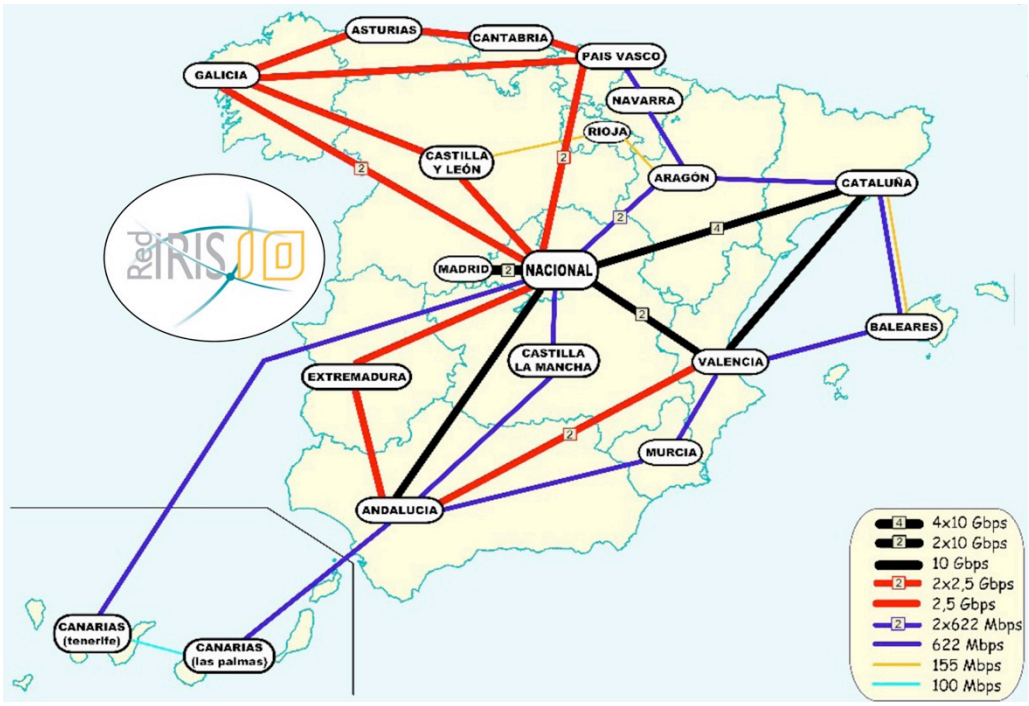

Figura 3-1: Topología RedIRIS-10, 2007-2011 [30]

3.2 Monitorización basada en flujos

Dado el carácter principalmente académico del proyecto RedIRIS y de acuerdo a la ley española, los flujos de red se almacenan en servidores aislados y sin ningún tipo de tratamiento a nivel de flujo [14] [25] que revele la información del usuario; únicamente guardan datos relativos a las distintas conexiones. Esto facilita su posterior utilización según la finalidad para la que se empleen en cada momento. Asimismo, hay que tener en

cuenta la tasa de muestreo del sistema de monitorización, de uno por cada doscientos paquetes.

Dentro de los proyectos nacionales de I+D DIOR y ANFORA se desplegó un procedimiento de monitorización basado en NetFlow cuyas salidas utilizaremos en este trabajo. El proceso comienza con la recolección de los datos mediante Flow-Tools en un repositorio, del que se obtienen una serie de estadísticas gracias al sistema de procesamiento y que a su vez son almacenadas por el sistema de monitorización, donde se aglutinan los datos finales a los que tenemos acceso.

Más detalladamente, el sistema funciona de la siguiente manera: los datos binarios de cada exportador, es decir, los flujos exportados en binario, llegan mediante datagramas UDP al colector situado en la Universidad Autónoma de Madrid, donde se comprimen y agregan cada 15 minutos; el sistema descomprime cada día todos los ficheros, los filtra por exportador y extrae los flujos en un único fichero por día y exportador. En este punto, varias herramientas propias conforman los distintos ficheros de texto *finales*, de sencillo tratamiento e importación a un *script* programable.

3.2.1 Formato de los archivos de texto

El proceso de extracción con Flow-Tools permite almacenar la información de los flujos de red en ficheros de texto, de distintos tipos según los datos guardados en cada uno de ellos. Así, nos encontramos con ficheros .AB, .HC, .HH… Teniendo en cuenta el objetivo final de caracterizar los sesenta minutos del día con menor utilización, en este caso cobran especial interés, y por tanto serán la base del estudio, los ficheros .AB, siglas que se corresponden con *Ancho de Banda*. Estos archivos de texto presentan la siguiente nomenclatura: Exportador_añomesdia.txt.AB.

En los ficheros se encuentran catorce columnas, que se enuncian a continuación: tiempo unix, bytes medios, bytes máximos, nombre de la red, exportador, tipo de día, sentido del flujo, número de IPs activas en el origen, número de IPs activas en destino, número total de IPs activas en el origen, número total de IPs activas en el destino, fecha, fecha exacta y número de ceros en todo el día.

Como se ampliará en secciones sucesivas, adquieren gran importancia los datos de tiempo unix -segundos desde el 1 de enero de 1970-, bytes medios -bytes por segundo-, exportador, tipo de día -festivo o laborable, considerándose festivos los días entre semana fuera del calendario académico, como los de los meses de julio o agosto- y número de ceros -errores detectados ese día, en segundos-.

Cada línea del archivo de texto aglutina dichos parámetros cada cinco minutos para un exportador concreto, como se puede ver en la Figura 3-2.

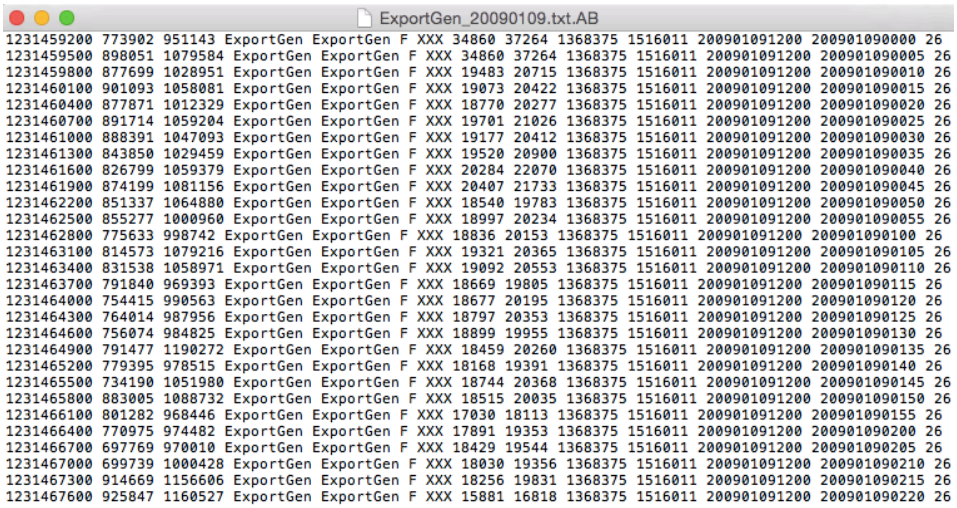

Figura 3-2: Ejemplo de fichero de texto .AB de un exportador genérico

3.3 Disponibilidad de las medidas

Los ficheros de texto sobre los que se sustenta este trabajo, extraídos según la metodología ya introducida en la Sección 2, empiezan el 1 de enero de 2008 a las 00.00h (GMT) y concluyen el 31 de diciembre de 2013 a las 23.59h (GMT). Sin embargo, se han encontrado algunas limitaciones en cuanto a la disponibilidad de datos que conducen al descarte de varios puntos de presencia en el análisis.

Con la migración comentada en la Subsección 3.1.2 de RedIRIS-10 a NOVA, la infraestructura de RedIRIS sufrió cambios importantes que llevaron a que, progresivamente, algunos de los exportadores salieran del programa de monitorización. Esto implica que varios de ellos sean descartados en esta fase inicial, pues se estima que sus estadísticas son poco representativas como para ser consideradas.

Tras la realización de una tabla que presenta con detalle los archivos de texto disponibles (expuesta brevemente en la Tabla 3-1 -las equis (X) corresponden a meses con al menos veinte días de datos; los guiones (-) , con meses con menos de cinco días-; completa en el Anexo A), se toma la decisión de prescindir de varios exportadores de cara a la realización del estudio. Así, Madrid7, Tenerife0 y Rioja0 son excluidos al no superar el año de muestras.

Año-mes/ Exporter	Badajoz0	Barcelona0	Bilbao0	CiudadReal0	LasPalmas0	Madrid0	Madrid5	Madrid7	Murcia0	Oviedo0	Palma0	Pamplona0	Rioja0	Santander0	Santiago0	Sevilla0	Tenerife0	Valencia0	Valladolid0	Zaragoza0
2008-01	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-02	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-03	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-04	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-05	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X

Tabla 3-1: Perspectiva de los archivos disponibles (ver Anexo A).

4 Análisis preliminar

La determinación del momento valle en diversos puntos de presencia de RedIRIS se inicia con un preanálisis centrado tanto en la función de distribución acumulada (FDA) como en su derivada, la función de densidad de probabilidad (FDP), aplicados a la hora menos cargada de cada exportador.

Antes de establecer las consideraciones definitivas que darán paso al modelado de cada exportador, se estudia el efecto que tiene uno de los parámetros de los archivos de texto enumerados en la Subsección 3.2.1, el denominado *número de ceros*, relativo a los errores -segundos con errores, más exactamente- que habido en la extracción de los flujos de red en un día concreto.

Además, se examinará la integridad de las medidas disponibles, pues estas abarcan seis años y múltiples PoPs. Estimamos que durante la monitorización hubo periodos de caída -en algunos casos, prolongados en el tiempo- y, consecuentemente, algunos PoPs denotaron fallos significativos; sin embargo, no se conoce ni está al alcance precisar con detalle estos errores, que podrían sesgar notablemente las conclusiones extraídas a partir de ellos.

En este punto del trabajo se hace igualmente necesario, para el uso posterior de ciertas herramientas de análisis, verificar la estacionariedad de la variable hora menos cargada y comprobar qué función pueden tener otros factores -como el ancho de banda o la influencia de la hora más cargada- en la hora de menor utilización de la red de cada PoP.

4.1 Visualización del fenómeno

4.1.1 Planteamiento

Tras prescindir de tres exportadores en la Subsección 3.3, los diecisiete restantes entran en la fase de preanálisis. Por cuestiones de privacidad de los datos y las subredes, en adelante se establece una sigla fija para cada uno de ellos, como PoP1, PoP2… PoP17.

La gran cantidad de datos disponibles hacen imprescindible una primera toma de decisión relativa al tratamiento de los mismos. Teniendo en mente la meta final del estudio y partiendo de la idea intuitiva de buscar un método que permita la comparación lo más inmediata posible entre distintos exportadores, se cree oportuno agrupar los seis años de estadísticas de cada PoP por separado. Este planteamiento se sustenta con la justificación de estacionariedad, detallada posteriormente en la Subsección 4.4.

Por otro lado, la búsqueda de la hora valle, conceptuada como los sesenta minutos consecutivos con menor ancho de banda a lo largo de un día, hace inevitable la toma de decisión sobre los flujos posteriores a las 23 horas. Un primer acercamiento al problema puede consistir en desechar estos, aunque dicha opción es rápidamente descartada por su notable alteración de los resultados. Finalmente, se decide *completar* las horas que comienzan entre las 23 y las 00 horas con los minutos necesarios del día siguiente. Esto conlleva que si el día consecutivo al que se está examinando presenta algún problema o no existe, el examinado se descarta.

Asimismo, antes de iniciar el análisis se hace notar que, por el hecho de agrupar datos cada cinco minutos, el ancho de banda en dicho periodo de tiempo se computa asumiendo que los bytes están distribuidos de manera uniforme. Esto acarrea unas imprecisiones en la estimación que pueden considerarse mínimas y no significativas frente al tamaño de una hora bajo estudio.

4.1.2 Definiciones

Para la observación eficaz de la hora valle, se ha encontrado necesario hacer uso de dos instrumentos matemáticos, la función de distribución acumulada y la función de densidad de probabilidad. Estas herramientas permiten construir gráficas con las que representar rápidamente el fenómeno.

Sea X una variable aleatoria y x cualquier número real en el rango [-∞,∞], la FDA F_x se define como [31]:

$$F_x(x) = P\{X \leq x\}, \text{ siendo } P\{X \leq x\} \text{ la probabilidad del suceso } \{X \leq x\}.$$

Por su parte, la FDP f_x se define como la derivada de la FDA F_x [31]:

$$f_x(x) = \frac{dF_x(x)}{dx}$$

Por su mayor simplicidad de cálculo, normalmente se optará por representar la FDA, aunque en algunas ocasiones se escogerá la FDP por su mayor conveniencia a la hora de comparar distribuciones y reflejar las medias y modas respectivas de cada una de ellas.

4.2 Integridad de los datos

4.2.1 Fuentes de riesgos

Antes de emitir cualquier juicio de valor o conclusión es necesario controlar y acotar las fuentes de riesgo sobre los datos finales que van a ser analizados. A lo largo de todo el proceso, desde que los flujos son extraídos por el sistema NetFlow en los exportadores hasta que se presentan las gráficas y resultados finales, existen varias amenazas de diversa índole para la integridad de las medidas. Estas se enuncian a continuación:

- **Cortes o interrupciones en la actividad normal de la red**. Aunque el sistema de monitorización funcione correctamente, el momento de la captura de los flujos puede coincidir con una caída de la red, ya sea a nivel local o global. En tal caso, podemos tratar con anchos de banda afectados por este problema, normalmente reflejando una actividad menor a la habitual.

- **Fallo en el envío de los flujos por los *routers***. Otra fuente de riesgo que nos encontramos es el fallo en el envío de los registros de los flujos de red desde los exportadores, por lo que nos topamos con datos incompletos.

- **Problemas relacionados con hardware.** Este riesgo puede aparecer cuando la monitorización y el envío de los flujos se llevan a cabo sin impedimentos, pero entran en acción contratiempos a nivel hardware, por ejemplo a la hora de almacenar los mismos en discos duros.

- **Errores al analizar los flujos del colector.** El mecanismo de análisis y extracción de medidas elaboradas a partir de los flujos almacenados puede ser una fuente de problemas, producidos por los scripts encargados de la manipulación, ya sea por falta de sincronismo o de recursos ante el análisis de una situación anómala de tráfico [33]. Su solución pasa por la detección de los mismos localizando días incompletos y el relanzado de los programas para la recuperación de los archivos.

- **Diacronía a medianoche.** Se distinguen dos posibles amenazas para la integridad de los datos que tienen que ver con las 00 horas. Por un lado, según lo indicado en la Subsección 4.1.1, unir las estadísticas desde las 23 horas con las del día siguiente propicia injerencias en futuros análisis de factores como si el día es laboral o festivo; por otro, se han apreciado singularidades con los primeros datos del día que hacen pensar en oscilaciones que contaminan los primeros minutos.

- **Tareas de *backup* propias de cada subred que afectan al propio análisis.** En algunos exportadores se han revelado comportamientos extraños, en forma de ocupaciones altas a horas del día con, intuitivamente, escasa actividad. Esto nos guía a deducir que en algunos PoPs se efectúan labores de tipo *backup* que deben ser tenidas en consideración para que no influyan en nuestras propias conclusiones.

4.2.2 Tolerancia a errores: comparativa

Relacionado con lo expuesto en la Subsección 4.2.1, uno de los útiles que nos permiten delimitar los errores es el número de ceros, a los que nos referiremos también como fronteras de error (FE). De los 86.400 segundos que hay en un día, esta columna de los archivos de texto refleja qué segundos han sido cero; es decir, sirve para contabilizar las muestras que se pueden estimar como erróneas.

En la Figura 4-1 se puede comprobar que el 95% de los ficheros disponibles cuentan con un número menor o igual a 1075 ceros y el 80%, menor o igual a 52 ceros.

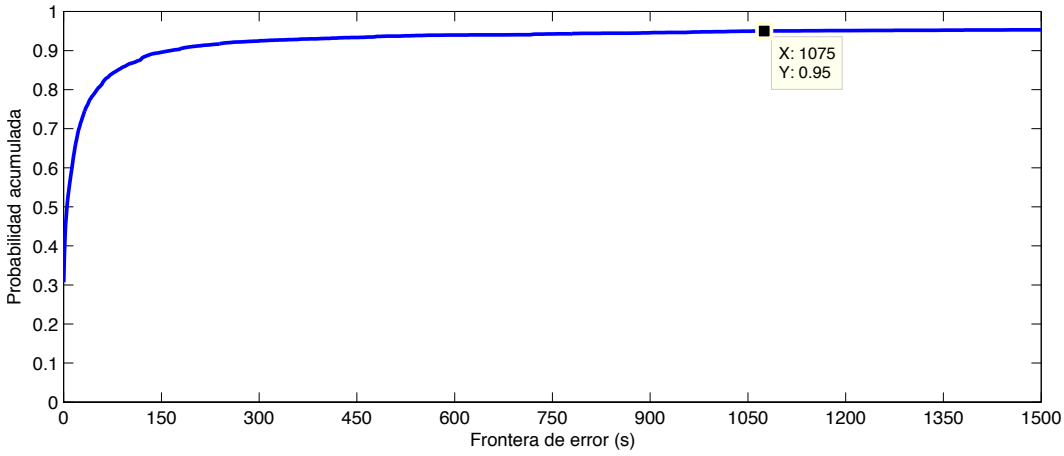

Figura 4-1: Distribución acumulada de las fronteras de errores

Las FE se tratan de la siguiente manera: se establece un límite superior por el cual, si el exportador bajo estudio cuenta con un número de ceros mayor para ese día, ese día se suprime y sus características no entran en el estudio. Este punto también atañe al día siguiente al analizado -similar a lo expuesto en la Subsección 4.1.1-.

Para efectuar una comparativa entre las distintas fronteras, se toman como muestras representativas del conjunto de los exportadores los casos de los PoP12 y PoP14, evaluados con FE 1, 5, 25, 50, 100 y 1000. Dichos PoPs encarnan los dos patrones observados entre los exportadores: los que, como PoP14, apenas presentan variación en su FDA pese a admitir más errores, y los que, análogos a PoP12, son sensibles a las FE.

En la Figura 4-2 se presentan las FDA de PoP14. En el título de cada gráfica se señalan la FE y los días válidos totales de ese PoP. Lógicamente, a mayor aceptación de número de ceros, más días se toman como válidos y, por tanto, más muestras conforman el análisis e intuitivamente obtenemos resultados más significativos.

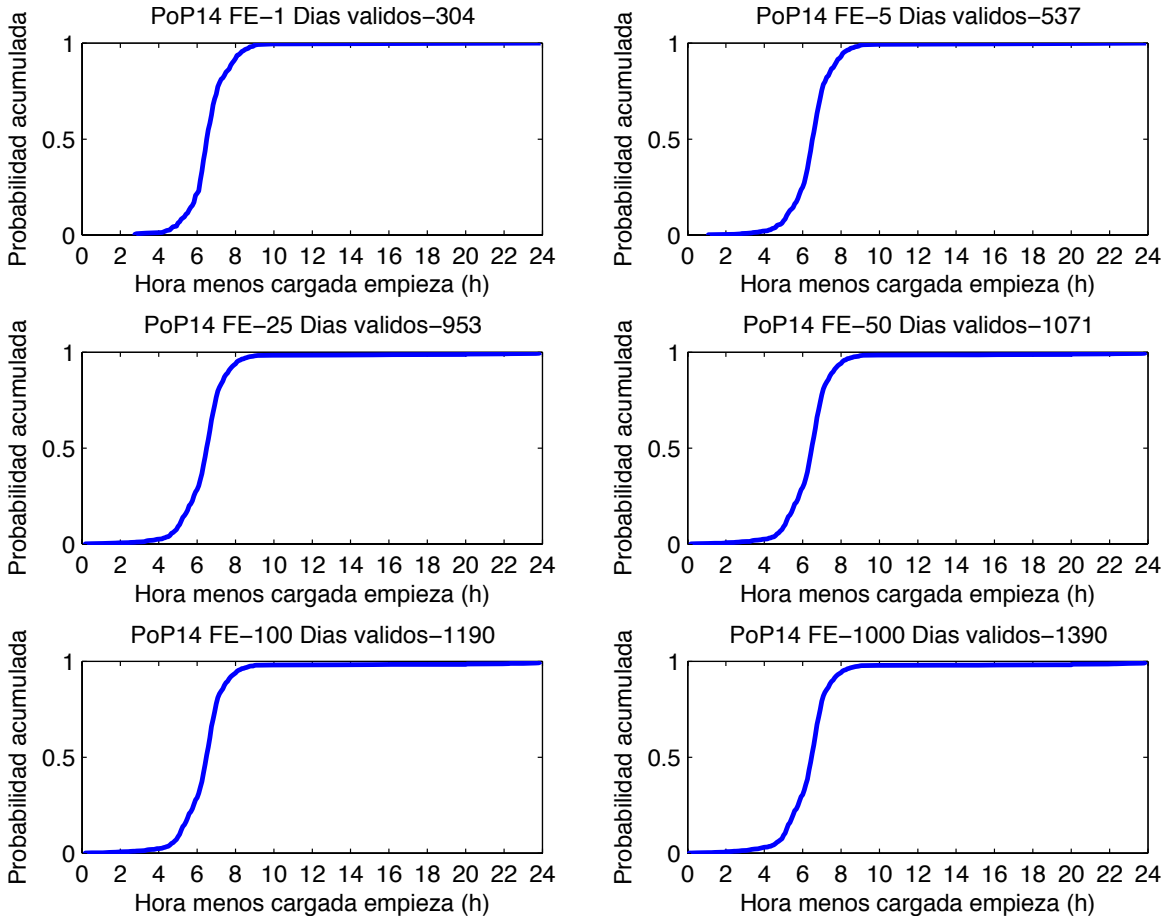

Figura 4-2: Comparativa de fronteras de error para un mismo exportador (PoP14)

Se aprecia que la forma de la FDA del momento valle se mantiene muy similar pese al gran aumento de las muestras y no se vislumbra que los días con FE superiores perviertan el análisis. Esto nos lleva a elegir, para los exportadores fieles a este patrón, FE altas, con el objeto de enriquecer el estudio y abarcar el mayor número de días posibles.

Por su parte, en la Figura 4-3 se muestran las FDA correspondientes al PoP12, con la misma nomenclatura escogida para la Figura 4-2.

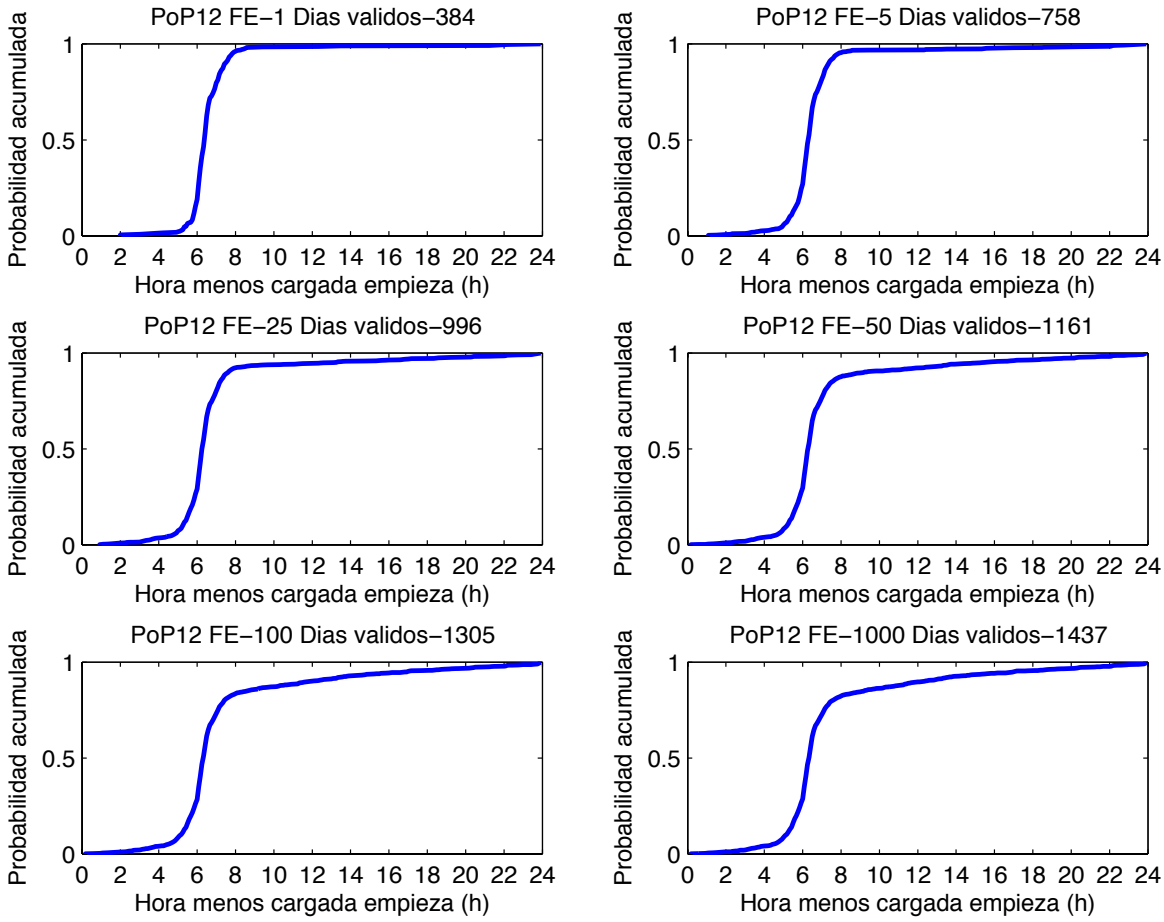

Figura 4-3: Comparativa de fronteras de error para un mismo exportador (PoP12)

En este exportador se distingue claramente un cambio notorio en la FDA según se aumenta la frontera. Observamos que, a FE mayores, el volumen en torno a las 9 horas va teniendo más peso en la distribución. Consecuentemente, valoramos que, en los PoP semejantes a PoP12, los días con mayor número de errores introducen muestras que contaminan el estudio, desplazando una parte de la masa de la hora a valle a instantes centrales del día.

4.3 Exportadores a estudio

4.3.1 Perturbaciones percibidas

Los riesgos señalados en el apartado 4.2.1 provocan que algunos PoPs presenten comportamientos no fácilmente justificables, vinculados a errores de distinta naturaleza. Estos exportadores serán apartados del estudio, puesto que sus inclusiones en el mismo desvirtuarían por completo el objeto del análisis.

El trabajo en esta fase se centra en la visualización de las FDA de todos los exportadores, con especial énfasis en la búsqueda de particularidades específicas de cada uno de ellos.

La figura 4-4 muestra la FDA de PoP15, uno de los *routers* en los que se han encontrado comportamientos singulares.

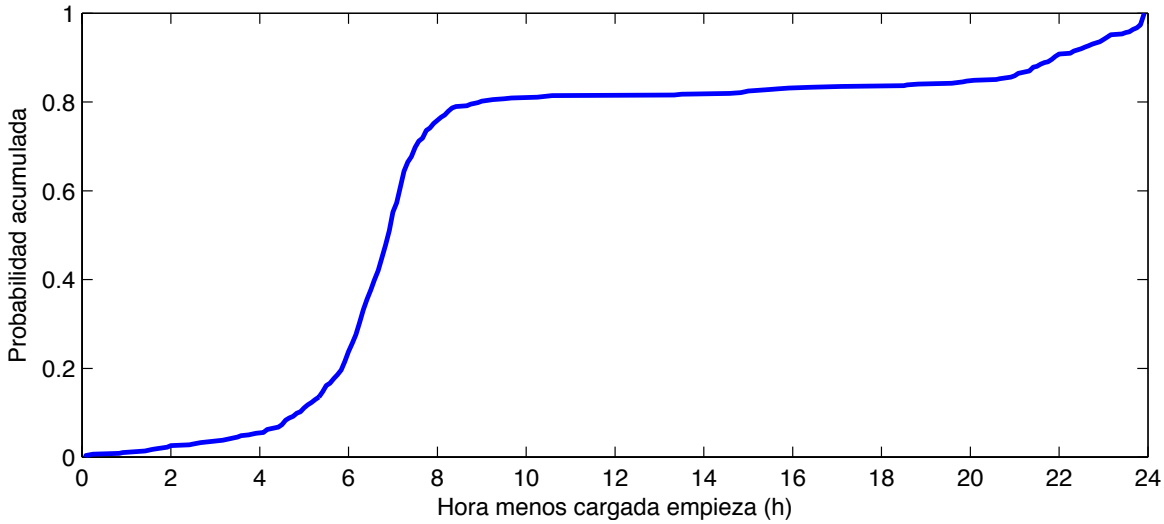

Figura 4-4: Distribución del momento valle en PoP15

Específicamente, se observa un 20% aproximado de horas con menor carga a lo largo del día a partir de las 10 horas. Además, principalmente por el repunte de la probabilidad a partir de las 20 horas, y de forma más llamativa, desde las 22 horas, concluimos que ese último lapso de tiempo se ve afectado por tareas de mantenimiento de la propia subred de cualquier género.

Otro de los exportadores en los que se han detectado excepcionalidades es PoP16, cuya FDA se muestra en la Figura 4-5.

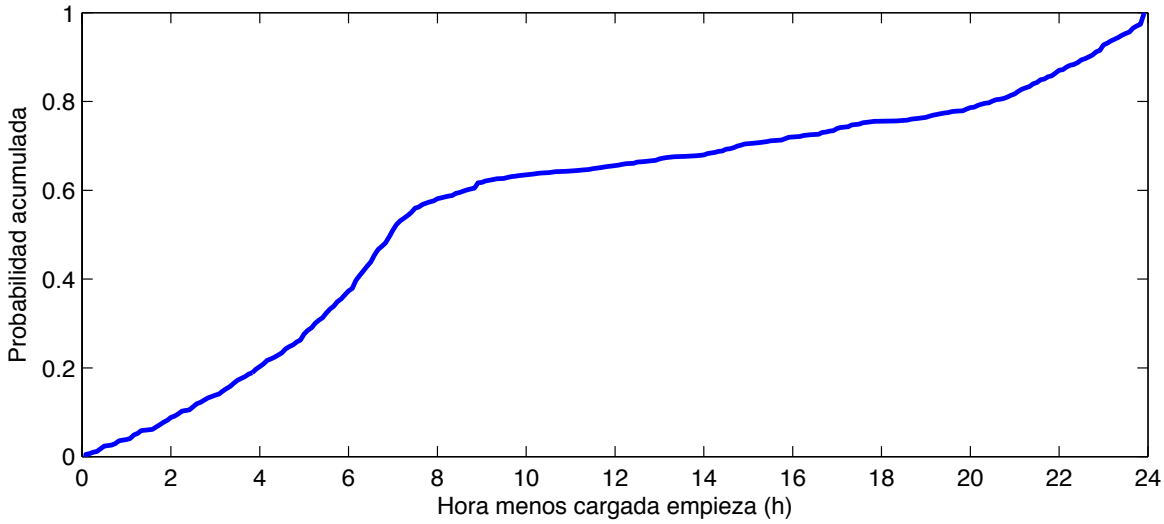

Figura 4-5: Distribución del momento valle en PoP16

Dada la tendencia aún más marcada en cuanto a la uniformidad temporal de las alteraciones, para este exportador en concreto se desarrolla un examen más profundo, consistente en plantear por separado las distribuciones de probabilidad para días laborables y días festivos, puesto que una de las hipótesis recurrentes es una fuerte

variabilidad entre ambos tipos de día. Esta evaluación del factor día laborable o festivo será una de las claves de la predicción del momento valle en la Sección 6.

En la Figura 4-6 se muestran ambas FDA para PoP16, con una misma FE de 100.

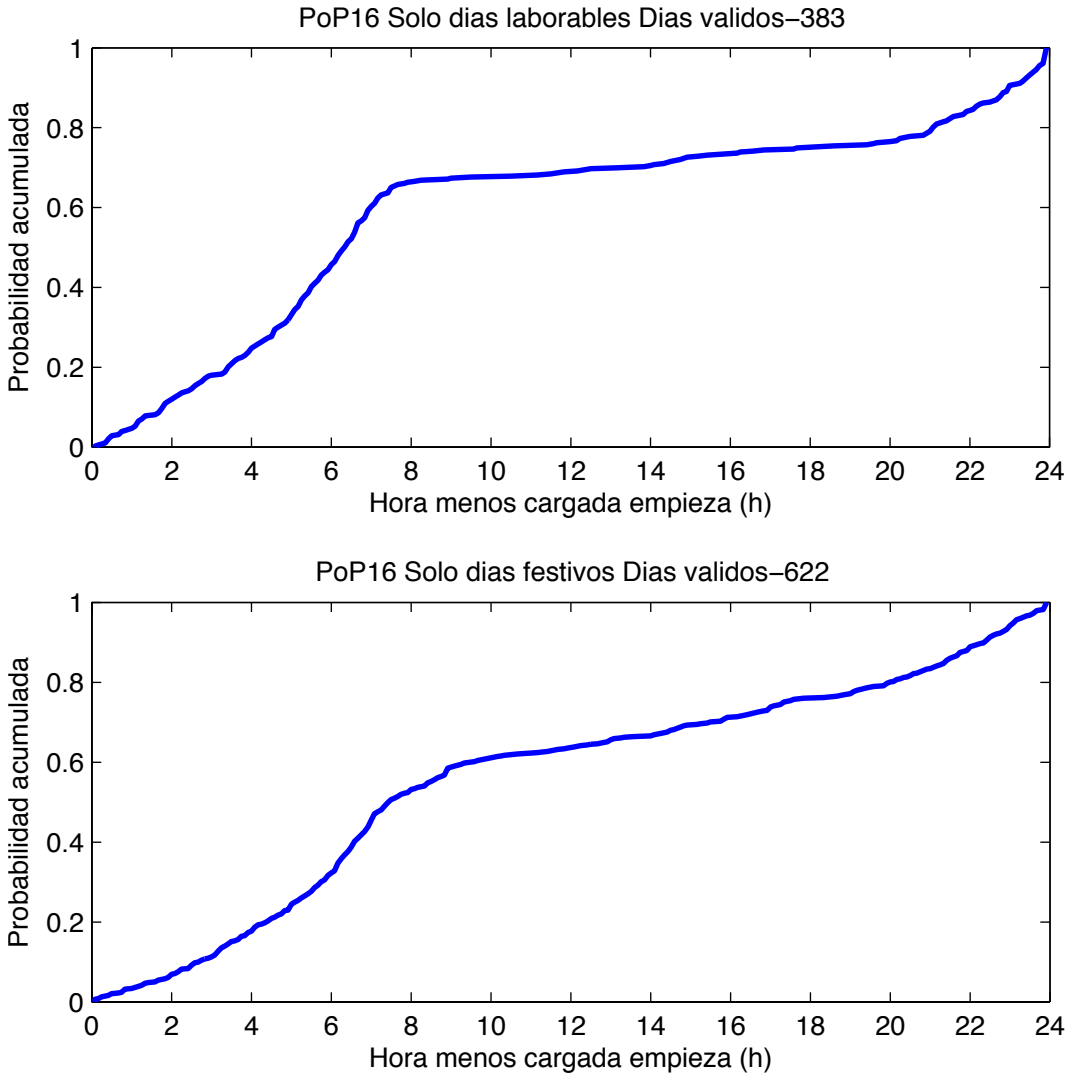

Figura 4-6: Distribución del momento valle en PoP16; L/F por separado

Efectivamente, la idea preconcebida es acertada y los días festivos siguen una pauta que tiende a ser uniforme; aún así, los días laborables presentan el comportamiento descrito para PoP15 en la Figura 4-4. Por tanto, ambas distribuciones se ven afectadas por alguno de los riesgos para la integridad de los datos, luego esta subred es descartada.

Sin embargo, debido al propósito de reunir el máximo número de medidas válidas a nuestra disposición, algunas peculiaridades se juzgan como tolerables por su baja afectación frente a la cuantía total de días válidos.

En la Figura 4-7 se muestran las FDA de PoP8 y PoP10.

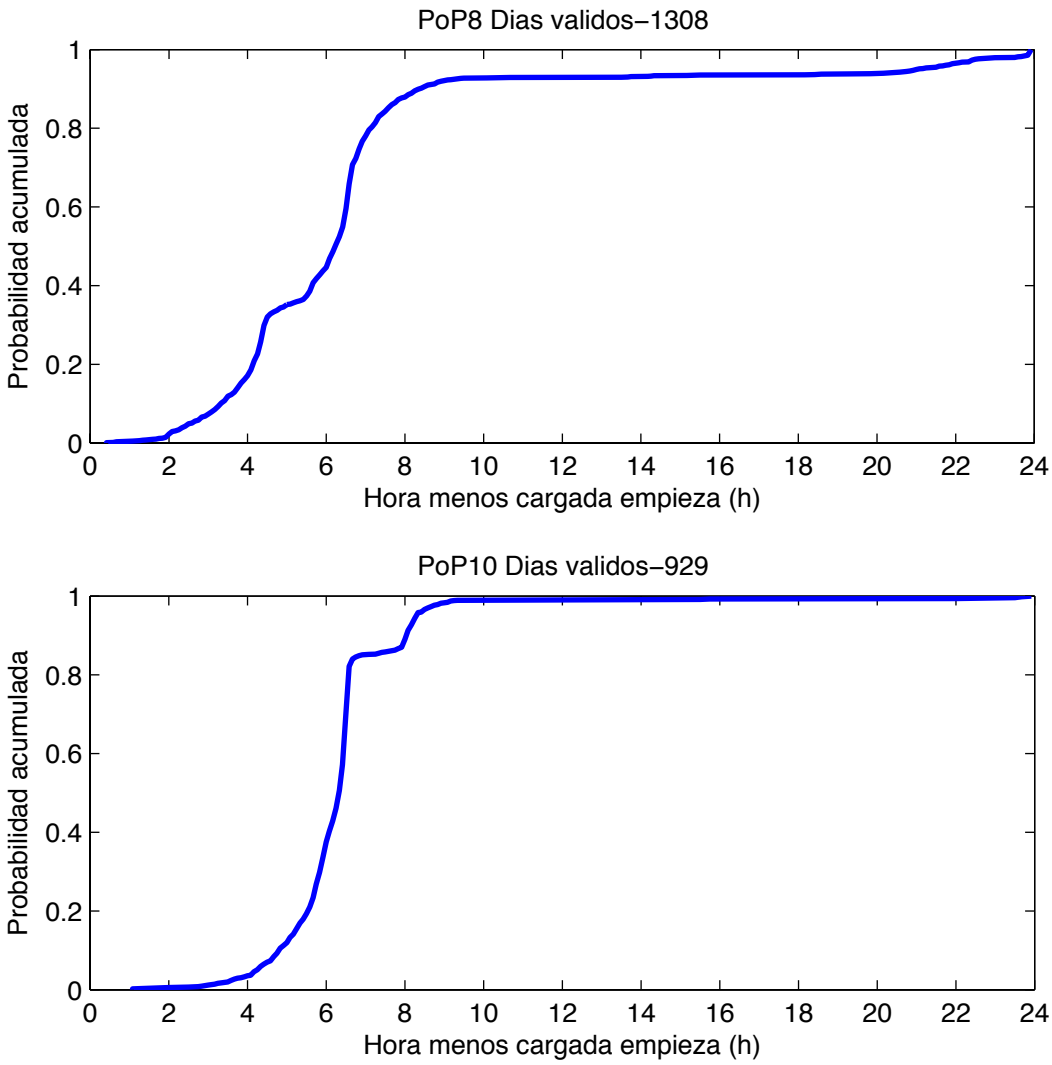

Figura 4-7: Distribuciones del momento valle en PoP8 y PoP10

PoP8 cuenta con una zona en torno a las 5 horas con baja probabilidad, presumiblemente por alguna labor de *backup -bulk-* de la subred, y varias horas valle posteriores a las 20 horas. Pero al contrario que para PoP15, en esta ocasión los consideramos aceptables por su escasa incidencia en la distribución. Igual premisa se sigue con el PoP10, con escaso volumen cerca de las 7 horas.

Como conclusión, las excepciones descritas en esta subsección y evaluadas para todo el grueso de los PoPs suponen que PoP15, PoP16 y PoP17 no prosigan dentro del análisis.

4.3.2 Determinación de los exportadores

Por todo lo comentado en la Subsección 4.3.1, finalmente nos encontramos con catorce exportadores que consideramos aptos para entrar a formar parte de un análisis significativo. El siguiente paso es determinar qué FE asignar a cada PoP. Esta elección se realiza con una doble intención: primordialmente, encontrar una relación entre la FE y número de días válidos que sea conveniente -dar prioridad al número de muestras sin introducir errores indeseables-; al mismo tiempo, declinarse por un tamaño de muestras

similar entre exportadores para que la trascendencia de cada uno de ellos en el análisis global sea lo más equitativa posible.

En la tabla Tabla 4-1 se especifican los exportadores, las FE por las que se ha optado y sus días válidos.

Exportador	FE	Días válidos
PoP1	100	946
PoP2	100	962
PoP3	100	1002
PoP4	1000	480
PoP5	100	1067
PoP6	100	992
PoP7	100	970
PoP8	100	1038
PoP9	100	906
PoP10	100	929
PoP11	50	1204
PoP12	5	758
PoP13	25	1279
PoP14	50	1071

Tabla 4-1: Exportadores, FE definitivos y días válidos

4.4 Estacionariedad

4.4.1 Explicación

En el contexto de un análisis que abarca un periodo extenso de tiempo, seis años, es ineludible dictaminar la estacionariedad de los datos. El concepto de estacionariedad adquiere especial importancia en este caso, pues de concluir que estamos ante un problema que no varía con el tiempo, la variable temporal puede ser apartada totalmente de la predicción del momento con menor utilización.

Una variable estacionaria no se ve afectada por cambios de origen temporal, o lo que es lo mismo, su distribución es finita y no varía al realizar un desplazamiento en el tiempo [32]. Sin entrar en definiciones matemáticas concretas e intuitivamente, podríamos decir que una variable es estacionaria si no tiene tendencia y sus muestras no se alejan de la media total a partir de un punto temporal concreto.

Por simplicidad, en este trabajo se analizará la estacionariedad como prueba visual.

4.4.2 Comprobación

La Figura 4-8 muestra la relación entre el momento valle y la variable temporal para todo el conjunto de los exportadores.

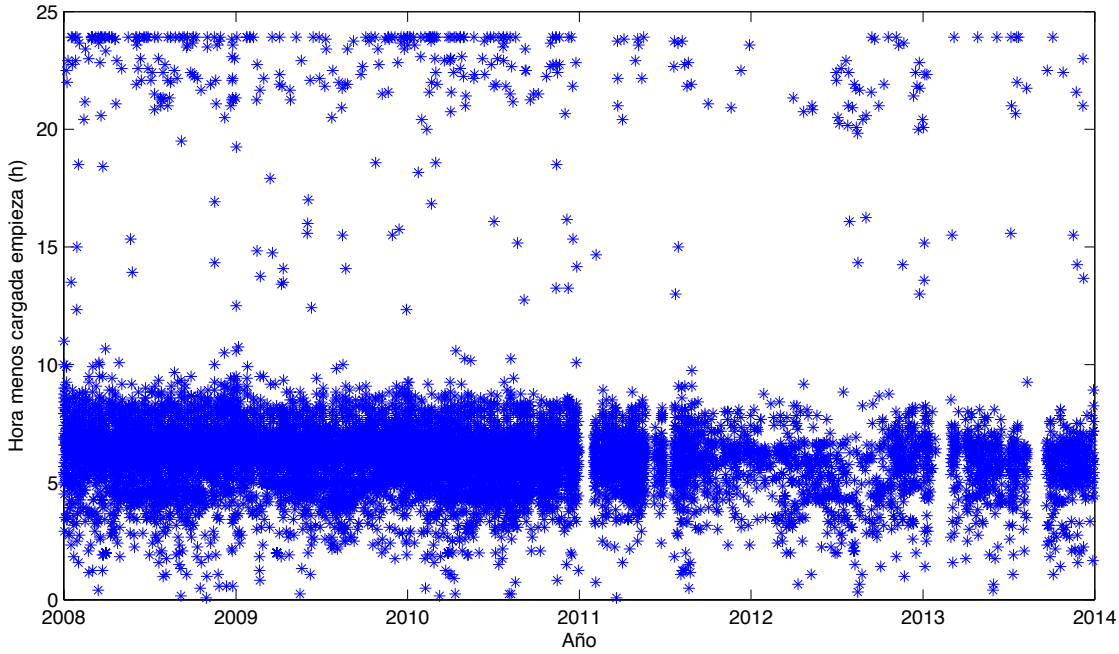

Figura 4-8: Representación temporal del momento valle

Observamos que el reparto de las muestras no presenta tendencias marcadas con el tiempo; tanto en el primer año como en el último la mayor parte de la masa está entre las 3 y las 9 horas y fuera de ese rango el número de datos no sufre grandes variaciones.

No obstante, para una mejor visualización del fenómeno, en la Figura 4-9 se representa la misma gráfica pero referida a un único exportador, PoP13.

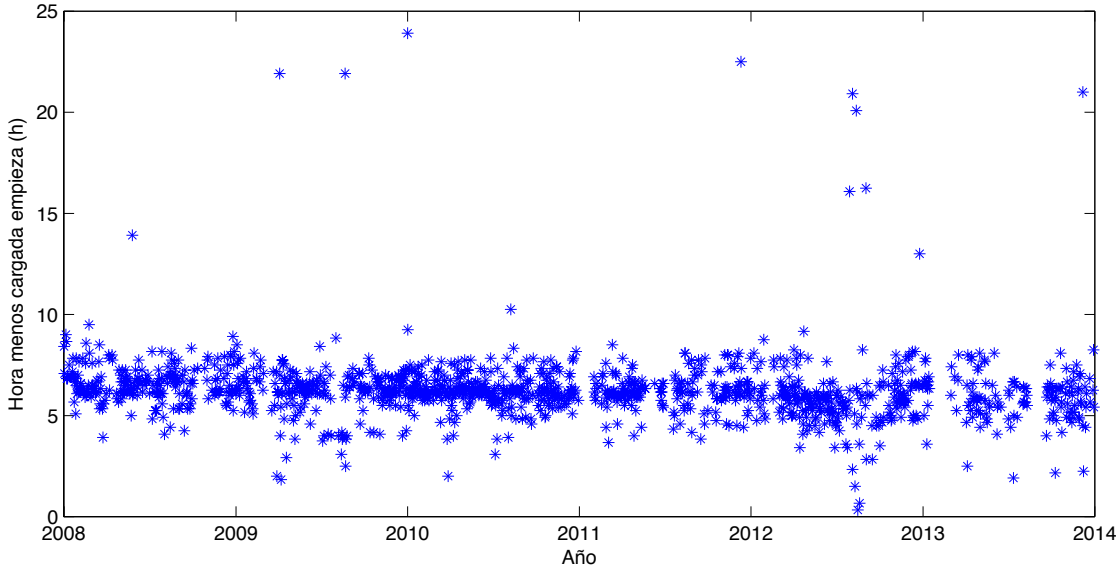

Figura 4-9: Representación temporal del momento valle para PoP13

Centrando la figura en un único exportador es aún más palpable la ausencia de tendencias evidentes, lo que nos lleva a concluir que podemos considerar que nuestro problema es estacionario.

4.5 Factores externos

4.5.1 Introducción

Además de la estacionariedad de la variable momento hora menos cargada, otra comprobación necesaria es juzgar el papel que puedan ejercer factores externos en cuanto a su afectación al problema en discusión.

Como se justifica en la Subsección 4.4.2, el momento de la hora valle es independiente de la variable tiempo. Sin embargo, otros factores o medidas complementarios al que nos ocupa sí que aducen cambios. Por ejemplo el ancho de banda a lo largo del día ha mostrado clara falta de estacionariedad.

En la Figura 4-10 se muestra esta característica para PoP8. Varias son las razones posibles de esto; conociendo la infraestructura concreta opinamos que es consecuencia de un cambio en el muestreo del sistema de monitorización, pero podría ser también fruto del incremento del uso -cambio de políticas de filtrado o aparición de alguna aplicación de relevancia- o del número de usuarios de Internet en los PoPs.

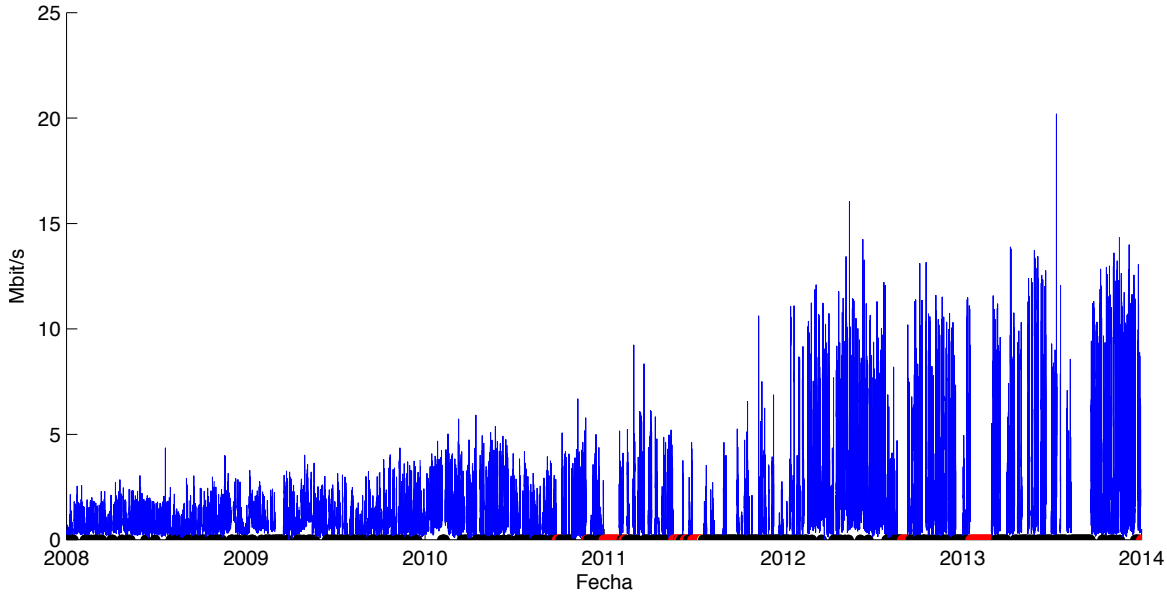

Figura 4-10: Ancho de banda en PoP8

Observamos que a partir de 2012 el ancho de banda medio durante todo el día es notablemente mayor, pasando de 2 Mb/s a 7 Mb/s aproximadamente.

4.5.2 Ancho de banda en el momento valle

La tendencia detectada en el Apartado 4.5.1 se ve reflejada en el valor del ancho de banda del momento de menor utilización, en el que efectivamente no se puede dejar a un lado la variable tiempo, como se representa en la Figura 4-11.

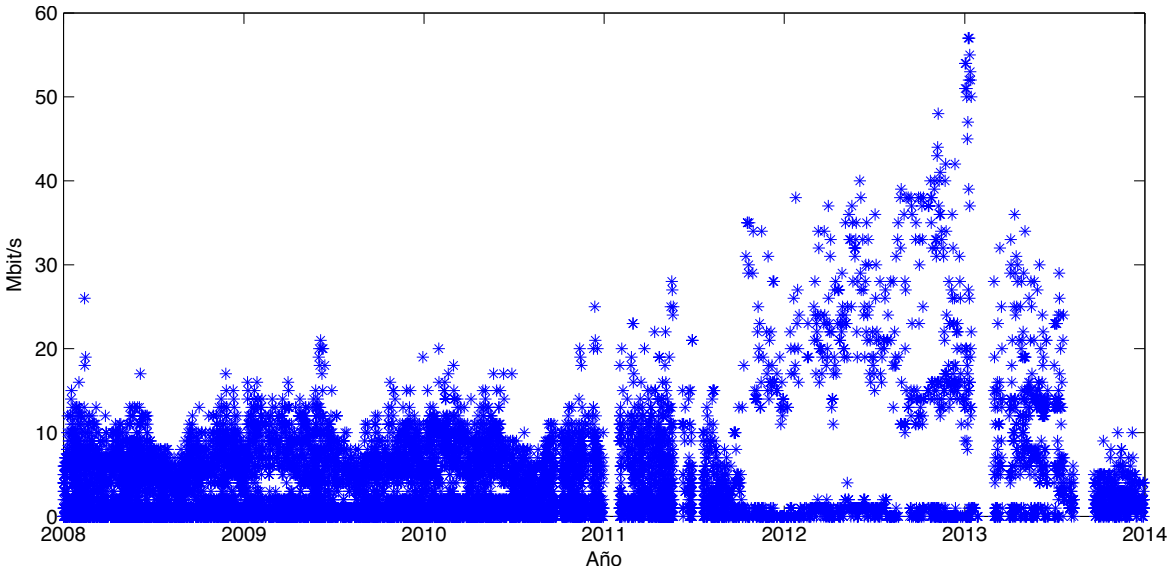

Figura 4-11: Representación temporal del ancho de banda del momento valle

Pese a ello, el interés radica en constatar su posible efecto en el estudio. En la Figura 4-12 se muestra cómo se reparten los anchos de banda medios en la hora valle y el momento en el que comienza dicha hora valle.

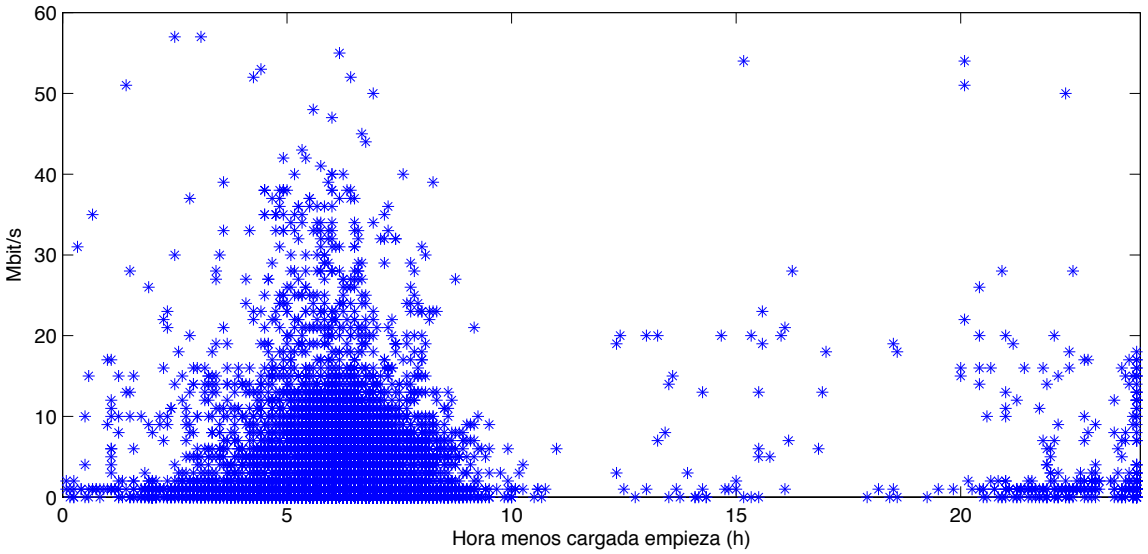

Figura 4-12: Ancho de banda frente al momento valle

Se percibe que los valores medios de ancho de banda superiores a 30 Mb/s, que según la Figura 4-11 están condicionados íntegramente por la variable temporal, no siguen una distribución especialmente diferente al resto de muestras, por lo que no se estima un factor significativo en nuestro análisis.

Para dotar de más solidez a esta afirmación, se opta por introducir el concepto del coeficiente de correlación, una medida del grado de relación entre las dos variables cuantitativas [34]. El cálculo de este coeficiente, de valor calculado 0'0192, confirma que es razonable afirmar ambos factores no están correlados y, por tanto, no tienen influencia uno en el otro.

4.5.3 Momento hora más cargada

Otro posible factor externo a tener en cuenta es la hora de mayor utilización y su papel en el inicio del momento valle. En la Figura 4-13 se muestra la relación entre ambos.

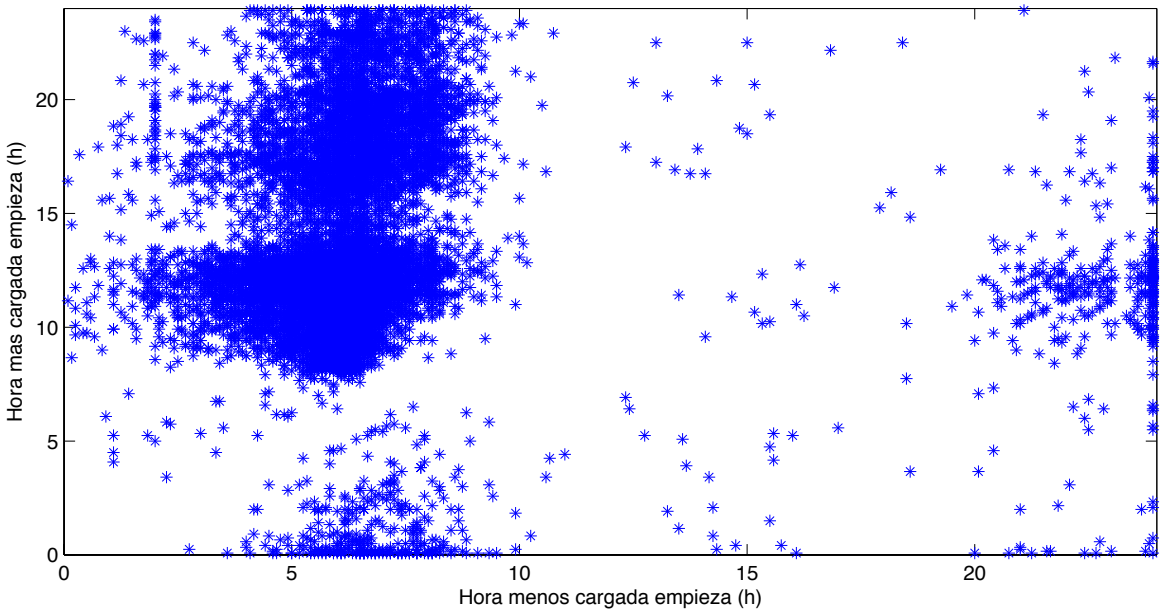

Figura 4-13: Momento hora más cargada frente a momento hora menos cargada

La interpretación de la Figura 4-13 no parece sencilla ni directa, por lo que se decide introducir de nuevo el coeficiente de correlación, con el que se obtiene un valor de -0’0095.

El signo negativo implica que se establece una relación inversa entre ambas, pero el valor es tan bajo que podemos concluir que las dos variables no están correladas y por tanto no nos servirá calcular la hora con mayor utilización como un indicio de cuándo ocurrirá la hora menos cargada.

4.5.4 Conclusiones

Los factores externos estudiados han ahondado en el preanálisis del momento valle. Por su posible influencia en los resultados finales, adquiere especial relevancia profundizar en la predecible no estacionariedad del ancho de banda en los exportadores, si bien se ha visualizado y evaluado que esta particularidad no provoca una alteración manifiesta en el instante en el que empieza la hora menos cargada.

De manera similar, se ha comprobado que el inicio de la hora con más carga en un día influye marginalmente en el momento de menor utilización. Especulábamos con que, si el lapso de tiempo en el que las subredes están más sobrecargadas se desplaza a un horario del día más tardío, esto provocaría que la hora valle se desplazara a un horario más temprano ese día, pero no hemos encontrado indicios de este comportamiento.

5 Caracterización

Uno de los principales objetivos propuestos en este trabajo es la búsqueda de la hora de menor utilización de una subred o un grupo de subredes dentro de RedIRIS. Su deducción y predicción exige, además de la evaluación de los factores que la condicionan y potencialmente explican, la propuesta de un modelo que caracterice la distribución típica del momento valle, estimando la proximidad entre distintos modelos teóricos y los datos empíricos con los que contamos. El objetivo es, primero comparar y describir las similitudes y diferencias entre distintas poblaciones o redes, y luego aplicar este conocimiento para emitir un pronóstico lo más certero y riguroso posible.

La propuesta se encamina mediante dos procedimientos paralelos. Por una parte, un modelado normal o gaussiano, vía natural después de examinar visualmente en la Sección 4 las funciones de distribución acumulada. Por otra parte, un modelado uniforme, pues se entrevé una considerable analogía entre éste y la distribución experimental en la región de mayor densidad de probabilidad. Estas proposiciones se analizan sin descartar una transformación de las muestras que permita un mejor ajuste a cualquiera de ellas.

La comparativa entre las dos alternativas establece cuál debe tomarse como más conveniente para nuestro trabajo, las razones que llevan a ello, las puntualizaciones necesarias para tomarla como válida y los efectos sobre el estudio.

5.1 Aproximaciones

5.1.1 Distribución normal

Según las primeras gráficas de la hora valle expuestas en la Sección 4 -como en la Figura 4-2-, una rápida reflexión -claramente se aprecia una única moda a partir de la cual y hacia ambos extremos la probabilidad cae- dirige a proponer un modelado normal o gaussiano para nuestro problema. La distribución normal o gaussiana es la más estudiada entre las FDP más populares y se utiliza en prácticamente todas las áreas de las ciencias e ingeniería porque describe de forma precisa muchas magnitudes de la naturaleza, en especial cuando éstas son producto de pequeños efectos/errores aleatorios o suma de varios procesos.

Una FDP normal se define como [31]:

$$f_x(x) = \frac{1}{\sqrt{2\pi{\sigma_x}^2}} e^{\frac{-(x-a_x)^2}{2{\sigma_x}^2}}$$, siendo a_x una constante y σ_x positivo.

En ella cobra especial interés la varianza, ${\sigma_x}^2$, ya que junto a la media μ -de la que es simétrica, e igual a la moda y la mediana- permite especificar inequívocamente una función de densidad normal. También permite establecer intervalos de confianza de especial utilidad, pues a la distancia de la media de *un sigma* -en el intervalo [μ - σ_x, μ + σ_x]- se aglutina el 68'26% de la probabilidad; a *dos sigmas* -en el intervalo [μ - $2\sigma_x$, μ + $2\sigma_x$]- el 95'44% y a *tres sigmas* -en el intervalo [μ -$3\sigma_x$, μ + $3\sigma_x$]- el 99'74% [35].

La Figura 5-1 muestra las FDP y FDA correspondientes a una variable normal, para diferentes medias y varianzas.

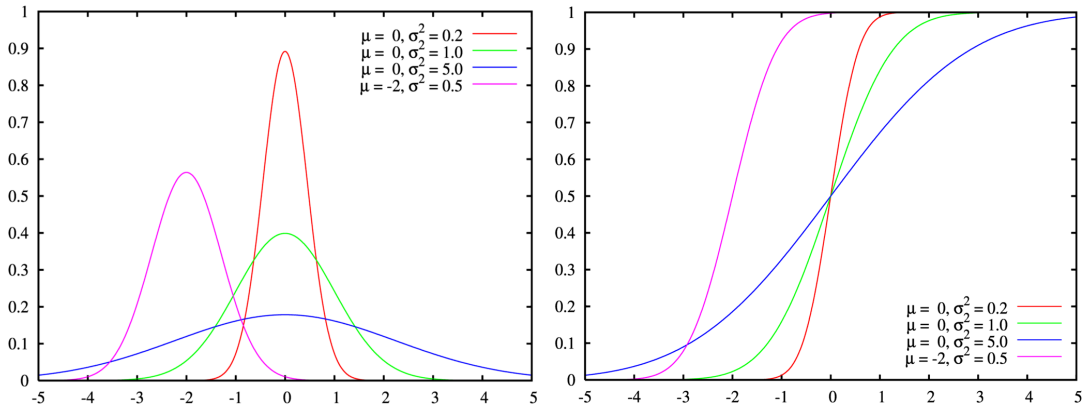

Figura 5-1: FDP (izquierda) y FDA (derecha) de una variable gaussiana genérica [36]

5.1.2 Distribución uniforme

La pendiente visualmente constante -o al menos próxima a ello- en las FDAs mostradas en la Sección 4, especialmente en la zona de mayor masa de probabilidad, hacen intuitivo pensar en un modelado uniforme continuo.

Una distribución uniforme se caracteriza porque todos los intervalos de la misma longitud dentro del rango [a, b] -ambos reales, siendo a una constante y b > a- tienen la misma probabilidad, o dicho de otra forma, la probabilidad de un suceso es independiente del punto entre a y b en el que se encuentre, sólo depende de los valores de a y b.

Una FDP uniforme se define como [31]:

$$f_x(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0 \quad en\ el\ resto \end{cases}$$

Por tanto, con a y b queda totalmente definida una FDP uniforme. A mayor distancia entre a y b, menor será la probabilidad para cada suceso.

La Figura 5-2 muestra las FDP y FDA correspondientes a una variable continua.

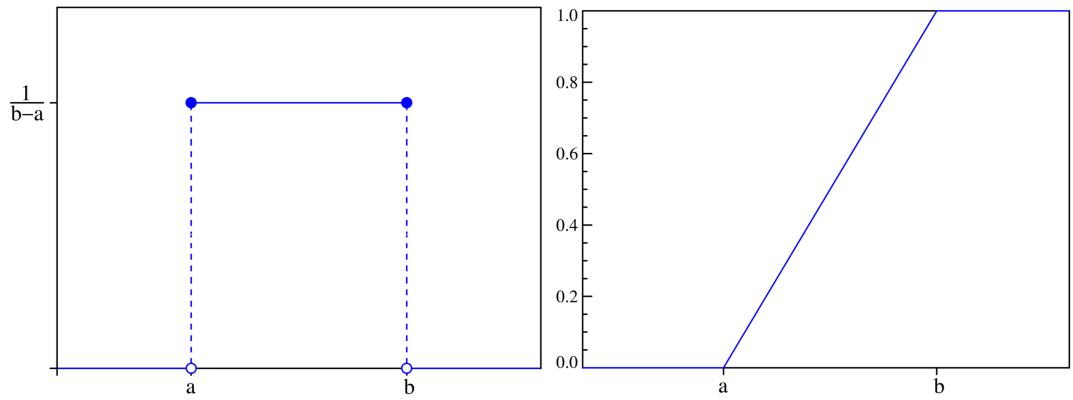

Figura 5-2: FDP (izq.) y FDA (der.) de una variable uniforme continua genérica [37]

5.2 Modelado normal

Como preámbulo de la propuesta de modelado normal o gaussiano, se asume la hipótesis nula de que la hora valle sigue una distribución normal con media μ y varianza σ^2. Esta hipótesis será validada o no según diferentes pruebas aplicadas a los datos.

5.2.1 Visualización mediante Q-Q Plot

Este primer método [38], cuya traducción literal al castellano sería *Gráfico Cuantil-Cuantil*, se basa en una comparación visual del grado de equiparación o ajuste entre cuantiles, puntos que dividen la función de la distribución en partes iguales. Su utilidad radica en que confronta las muestras empíricas con el valor de la función inversa a la FDA teórica, es decir, calculada a partir de la media y la varianza empíricas en el caso gaussiano.

La interpretación del gráfico es sencilla: donde la distribución empírica es idéntica a una distribución normal teórica, los puntos de ambos cuantiles se superponen.

La Figura 5-3 muestra el Q-Q Plot de los exportadores PoP4, PoP5, PoP6 y PoP7, tomados como representativos de la totalidad de los puntos de presencia. El comienzo de la hora valle, en el eje vertical, se expone en minutos desde la medianoche.

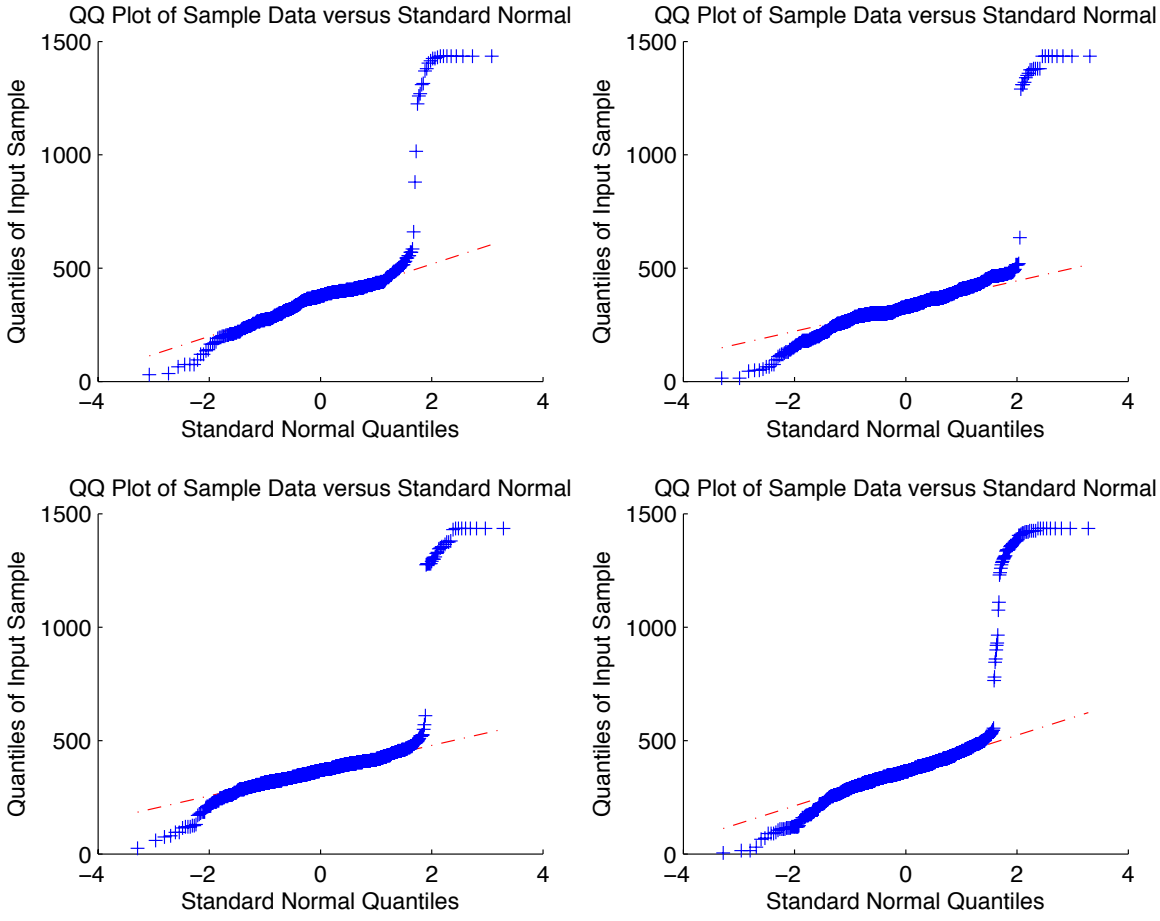

Figura 5-3: Q-Q Plot de PoP4 (sup. izq.), PoP5 (sup. der), PoP6 (inf. izq.) y PoP7 (inf. der.)

Aunque en distinta magnitud según cada exportador, hay una pauta que se repite en todas las representaciones: el modelado normal se ajusta correctamente en el centro de la distribución y moderadamente bien para momentos valle tempranos en el día -cercanos a las 00 horas por la derecha- pero no así para momentos valle tardíos -próximos a las 00 horas por la izquierda-.

En definitiva, el principal problema con el que nos encontramos es la presencia de más masa, esto es, más muestras tardías -más momentos valle que empiezan en las últimas horas del día- que las que tendría una distribución normal ideal teórica.

5.2.2 Test de Lilliefors

El test de Lilliefors es otro modo de probar que los datos introducidos en el test provienen de una variable distribuida normalmente. Fue especificado por Hubert W. Lilliefors [39] como una modificación del test no paramétrico de Kolmogorov-Smirnov para su aplicación sobre poblaciones de las que se desconoce la media y la varianza. Se trata de un test que precisa de datos muy robustos; presentimos que quizá excesivamente para un estudio con tantísima población como el nuestro.

Básicamente, el algoritmo encargado de dictaminar si se puede rechazar la hipótesis nula de normalidad de la distribución estima la media y la varianza a partir de los datos experimentales y calcula a partir de estas características una FDA, cuyos puntos son comparados mediante distancias con los de la FDA empírica. El test devuelve '1' si deniega la hipótesis nula para un nivel de significación del 5% -o lo que es lo mismo, si rechaza la distribución normal de nuestra variable hora valle-; '0' en el caso contrario.

En la Tabla 5-1 se muestra el resultado del test para cada exportador.

Exportador	Valor crítico	Valor estadístico	Lillietest
PoP1	0'0294	0'2915	1
PoP2	0'0291	0'2234	1
PoP3	0'0335	0'1377	1
PoP4	0'0411	0'3041	1
PoP5	0'0277	0'2218	1
PoP6	0'0287	0'2966	1
PoP7	0'0290	0'2773	1
PoP8	0'0250	0'2766	1
PoP9	0'0300	0'2205	1
PoP10	0'0296	0'2801	1
PoP11	0'0261	0'3153	1
PoP12	0'0328	0'2960	1
PoP13	0'0253	0'1550	1
PoP14	0'0276	0'2259	1

Tabla 5-1: Resultados del test de Lilliefors para cada PoP

Como los valores estadísticos calculados son en todos los casos superiores al valor crítico fijado por el test -dependiente del número de muestras-, se rechaza la normalidad de todos los exportadores.

5.2.3 Test de banda de ajuste sobre Q-Q Plot

Ante la no aceptación de la normalidad de nuestro problema por el test de Lilliefors, se decide incorporar un nuevo test de banda de ajuste basado en un test de correlación más indicado para estudios relativos a las redes de telecomunicaciones [40], con menores imposiciones, pero asumibles cuando se analizan numerosas medidas de tráfico. En esencia, esta prueba cuantifica el método visual descrito en la Subsección 5.2.1 empleando el coeficiente de correlación lineal entre el Q-Q Plot y las muestras empíricas.

Diversas publicaciones [41] justifican, apoyadas en el test de Kolmogorov-Smirnov para un nivel de significación del 5%, que la obtención de un coeficiente de correlación elevado -por ejemplo, muy próximo o superior a 0'9- nos permite concluir la normalidad de la distribución.

En la Tabla 5-2 se muestra el resultado del test para cada exportador.

Exportador	Test de banda de ajuste
PoP1	0'7076
PoP2	0'7629
PoP3	0'8337
PoP4	0'7518
PoP5	0'7213
PoP6	0'6819
PoP7	0'7607
PoP8	0'7994
PoP9	0'6992
PoP10	0'7594
PoP11	0'6572
PoP12	0'6747
PoP13	0'7959
PoP14	0'7223

Tabla 5-2: Resultados del test de banda de ajuste normal para cada PoP

El resultado del test para la mayoría de los exportadores, en torno a 0'7, no nos permite aseverar la normalidad de los datos.

5.2.4 Transformación de los datos

Una vez constatada la dificultad para certificar la distribución normal de la variable hora valle, se plantea realizar una transformación de los datos [42] con la que reafirmar la condición de distribución gaussiana.

Detectada en 5.2.1 la sospecha de que el problema mayoritario para la no normalidad se localiza en los momentos valle más retrasados en el día, se plantea una transformación de los datos, consistente en desplazar las horas valles más tardías al principio de la distribución como valores negativos para, en cierto modo, centrar la distribución en la media. Este tipo de transformaciones son un recurso habitual para conseguir normalidad sin alterar la integridad de las muestras.

Para encontrar el límite óptimo de la transformación se realizan varias simulaciones del test de banda de ajuste con los datos desplazados, siendo reflejados en la Tabla 5-3 los que han deparado resultados más idóneos.

Exportador	Desp. > 17h	Desp. > 18h	Desp. > 19h	Desp. > 20h
PoP1	0’8988	0’8988	0’8988	0’8955
PoP2	0’9617	0’9617	0’9617	0’9617
PoP3	0’9551	0’9551	0’9551	0’9551
PoP4	0’9235	0’9235	0’9235	0’9235
PoP5	0’9399	0’9399	0’9399	0’9399
PoP6	0’8620	0’8620	0’8620	0’8620
PoP7	0’9158	0’9164	0’9156	0’9156
PoP8	0’9432	0’9432	0’9427	0’9398
PoP9	0’9213	0’9213	0’9213	0’9213
PoP10	0’9188	0’9188	0’9188	0’9188
PoP11	0’8267	0’8267	0’8267	0’8267
PoP12	0’7866	0’7926	0’7916	0’7916
PoP13	0’9044	0’9044	0’9044	0’9044
PoP14	0’8841	0’8841	0’8841	0’8841

Tabla 5-3: Resultados del test de banda de ajuste normal para varios desplazamientos

Se aprecia una clara mejoría en el test respecto a 5.2.3 y que las diferencias entre los distintos traslados planteados son mínimas, si bien el que produce coeficientes de correlación más altos para todos los PoPs es el desplazamiento de los momentos valle que tienen lugar pasadas las 18 horas.

5.3 Modelado uniforme

El modelado uniforme surge como alternativa al modelado gaussiano por la similitud a éste de la distribución en la zona de mayor probabilidad. En este caso se escoge como test único el test de banda de ajuste presentado en la Subsección 5.2.3, modificado para amoldarlo como test de coeficiente de correlación entre los datos empíricos y la función inversa de una FDA uniforme, una recta.

5.3.1 Linealización a partir de todas las muestras

Una primera forma de afrontar el modelado uniforme es aproximando la distribución con la recta que mejor se acopla a la FDA de nuestro problema. La propuesta más inmediata y recurrente es un *curve fitting* o ajuste de curva con un polinomio de primer grado calculado a partir de todas las muestras de la distribución experimental.

En la Figura 5-4 se muestra un ejemplo de este procedimiento sobre la FDA de PoP13.

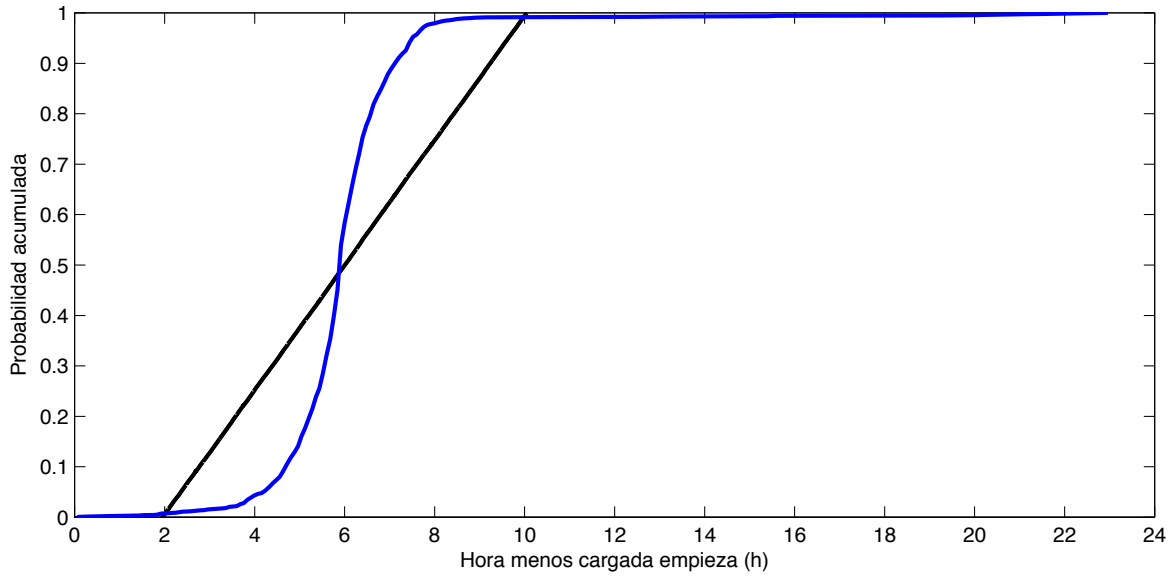

Figura 5-4: FDA correspondiente a PoP13 y su aproximación uniforme inicial

Detectamos dos inconvenientes en este modelo: el pobre ajuste de la recta calculada a partir de todas las muestras a la zona de mayor probabilidad y la gran separación entre la aproximación lineal y los momentos valle en las últimas horas del día. En la Tabla 5-4 se presenta el resultado del test de banda de ajuste uniforme para cada PoP.

Exportador	Test de banda de ajuste
PoP1	0’6113
PoP2	0’6870
PoP3	0’7573
PoP4	0’6625
PoP5	0’6249
PoP6	0’5802
PoP7	0’6709
PoP8	0’7279
PoP9	0’6013
PoP10	0’6637
PoP11	0’5475
PoP12	0’5625
PoP13	0’7035
PoP14	0’6216

Tabla 5-4: Resultados del test de banda de ajuste uniforme para cada PoP

Los valores ofrecidos por el test de banda de ajuste distan de ser aceptables para confirmar la uniformidad de la distribución de nuestro problema.

5.3.2 Linealización modificada

Con la finalidad de conseguir un modelado más apropiado y que revele coeficientes de correlación más elevados, se propone alterar el ajuste de curva realizado en 5.3.1.

Este nuevo planteamiento consiste en, a la hora de formular la recta, descartar los puntos fuera del rango aproximado donde se aglutina la mayor probabilidad, aproximadamente entre las 4 y las 9 horas. Así, se logra una recta que intuitivamente es más próxima a dicho rango, el de mayor relevancia en el análisis.

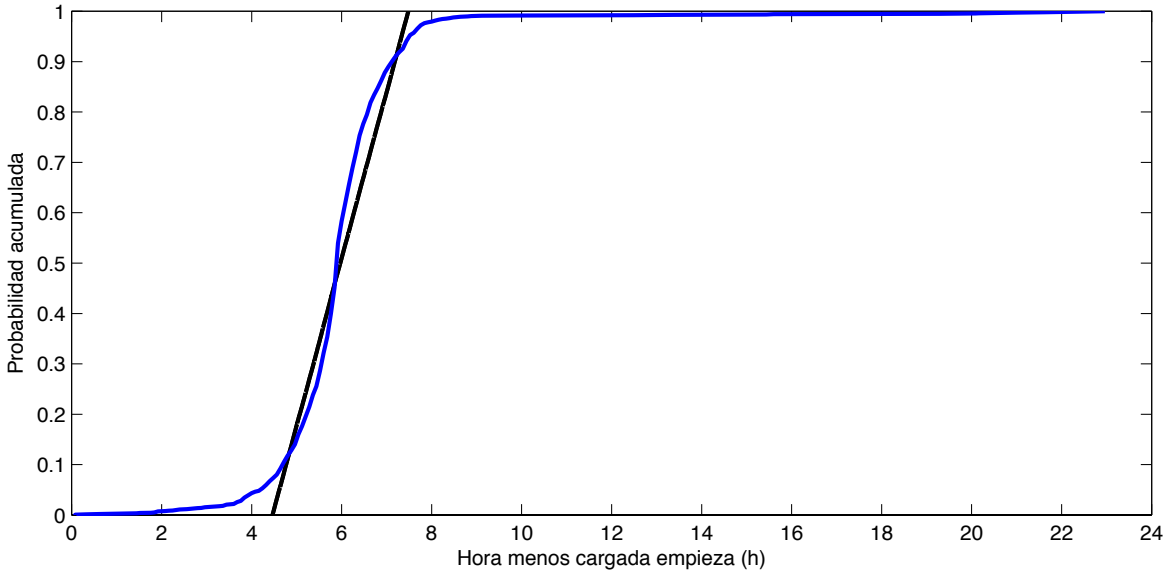

Figura 5-5: FDA correspondiente a PoP13 y su aproximación uniforme modificada

No obstante, y en contra a nuestro pensamiento inicial, en esta ocasión el test de banda de ajuste reporta exactamente los mismos resultados ya expuestos en la Tabla 5-4 para la linealización a partir de todas las muestras. Por lo tanto, esta variación del modelo no aporta ningún tipo de avance como se buscaba.

5.3.3 Transformación de los datos

Debido al *fracaso* con la linealización modificada, para el modelado uniforme se contempla igualmente la transformación por desplazamiento para las muestras posteriores a cierta hora, porque se ha comprobado su igual incidencia positiva en los resultados que la detallada en el caso normal en 5.2.4.

En la Tabla 5-5 se muestran los resultados del test de banda de ajuste con desplazamiento, fijado a las 18 horas -horas próximas apenas presentan cambios-, de nuevo erigido como el traslado óptimo.

Exportador	Test de banda de ajuste desplazado > 18h
PoP1	0’8171
PoP2	0’9157
PoP3	0’8961
PoP4	0’8543
PoP5	0’8761
PoP6	0’7829
PoP7	0’8379

PoP8	0’8898
PoP9	0’8499
PoP10	0’8359
PoP11	0’7220
PoP12	0’6735
PoP13	0’8200
PoP14	0’7978

Tabla 5-5: Resultados del test de banda de ajuste uniforme para varios desplazamientos

Al igual que en la caracterización normal, el desplazamiento hace que la consonancia entre las distribuciones experimentales de los PoPs y una distribución uniforme sea mucho mayor, superando para PoP2 el umbral de 0’9.

5.4 Comparativa y decisión

Como hemos visto, la *estrategia* de desplazar los momentos valle tardíos -después de las 18 horas- al principio de la distribución ha resultado satisfactoria tanto para el modelado normal como para el modelado uniforme. En la Tabla 5-6 se precisan los resultados del test de banda de ajuste para ambas suposiciones.

Exportador	Test normal	Test uniforme
PoP1	0’8988	0’8171
PoP2	0’9617	0’9157
PoP3	0’9551	0’8961
PoP4	0’9235	0’8543
PoP5	0’9399	0’8761
PoP6	0’8620	0’7829
PoP7	0’9164	0’8379
PoP8	0’9432	0’8898
PoP9	0’9213	0’8499
PoP10	0’9188	0’8359
PoP11	0’8267	0’7220
PoP12	0’7926	0’6735
PoP13	0’9044	0’8200
PoP14	0’8841	0’7978

Tabla 5-6: Comparativa del test de banda de ajuste para ambas distribuciones desplazadas

De manera manifiesta, el resultado del test normal es superior al del test uniforme, puesto que en todos los exportadores se obtiene un coeficiente de correlación sustancialmente mayor. Esto nos lleva a concluir que debemos modelar nuestra variable hora valle como una distribución normal.

Esta consideración queda formalmente respaldada por lo apuntado en la Subsección 5.2.3. Más del 64% de los coeficientes de los exportadores se encuentran por encima de 0’9; más del 85%, por encima de 0’86, y el valor más bajo obtenido roza el 0’8. Por consiguiente, consideramos que el modelado normal está suficientemente justificado.

5.5 Corolario

Descartados el uso de Q-Q Plot y el test de Lilliefors por su carácter excesivamente teórico y aceptada finalmente por el test de banda de ajuste la caracterización como distribución normal con desplazamiento, observamos el efecto que tiene este desplazamiento sobre la FDA, base del análisis.

Como ejemplo de la visualización, en la Figura 5-6 se muestra la FDA del exportador PoP7 con desplazamiento de muestras superpuesta a la FDA teórica calculada a partir de la media y la desviación típica. Esta y el resto de gráficas de los demás exportadores se encuentran disponibles en el Anexo B.

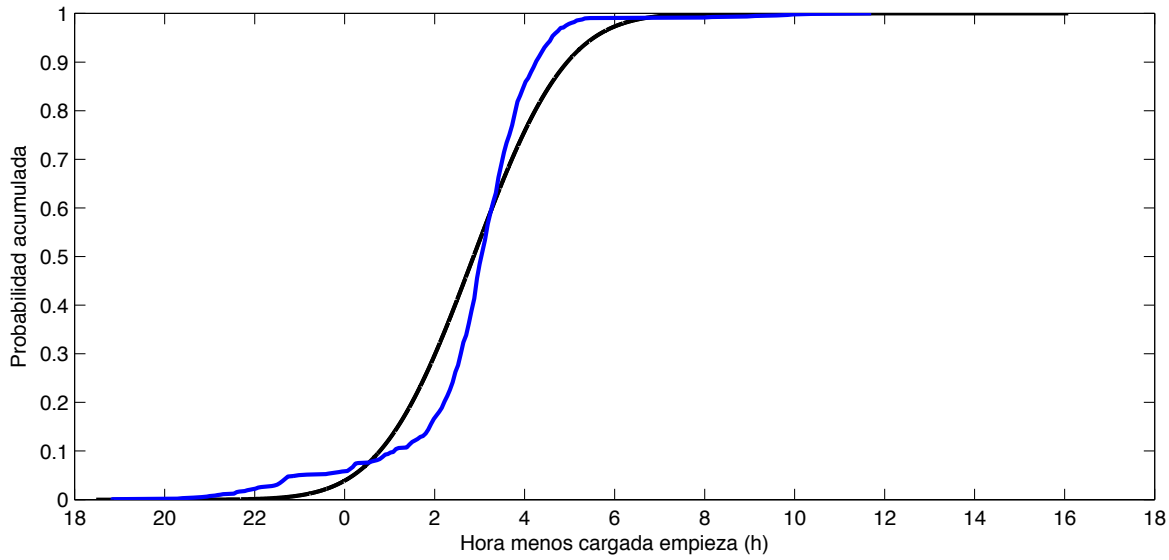

Figura 5-6: FDA con desplazamiento correspondiente a PoP7

La conclusión es que con el traslado de las horas valle que siguen a las 18 horas verdaderamente se consigue asemejar mucho más nuestra distribución a una distribución normal ideal, centrando la media en la distribución e igualando el tamaño de cada una de las *colas*, pasando a adquirir relieve en el grado de semejanza las pequeñas peculiaridades presentes en la distribución acumulada de cada PoP.

De lo expuesto, deducimos que los resultados del test de banda de ajuste normal con desplazamiento se ven amparados por el examen visual.

6 Predicción

La determinación y justificación de una caracterización normal para el problema de la hora valle da paso a la predicción de la misma, importante finalidad de este trabajo. Dicho modelado gaussiano, junto a propiedades ya examinadas como la estacionariedad del problema, nos permite acometer un procedimiento de análisis de la varianza (ANOVA) sobre los factores que se consideran de interés en este estudio, tales como el día de la semana, el mes, etcétera.

Pero antes de entrar de lleno en esta técnica, es razonable pararse a pensar cuáles son las peculiaridades en un análisis tan amplio como el nuestro, que pueden suponer resultados condicionados. Con la intención de ofrecer una solución lo más global y robusta posible, además del modelo propuesto por ANOVA se plantea un método alternativo, una aproximación fundamentada en los datos que disponemos que, ante la normalidad de la variable, defina un intervalo de confianza para cada uno de los exportadores.

Tras el análisis centrado en las subredes por separado, se conduce la predicción hacia la situación de hallar el momento ideal para realizar transferencias de tipo *bulk* en un enlace entre varios PoPs. Este problema será afrontado de varias formas, de acuerdo a las estrategias más inmediatas y también otras sugeridas en diferentes publicaciones relacionadas con este estudio.

6.1 Consideraciones previas

La formulación de un estudio que pretende englobar el máximo número de muestras, subredes y, en definitiva, datos, conlleva ciertas asunciones. Especialmente, se hace evidente el problema que supone la gran cantidad de elementos que componen el análisis, que pueden dificultar extraer conclusiones a través de modelos estadísticos como el que se empleará en esta sección, ANOVA.

Explicado y razonado en apartados anteriores, un examen visual por separado a las FDAs de cada PoP nos lleva a deducir la similitud en cuanto al momento valle en todos ellos. Para su mejor comprensión, en la Figura 6-1 se muestran las FDAs de todos los exportadores sobrepuestas en una única gráfica.

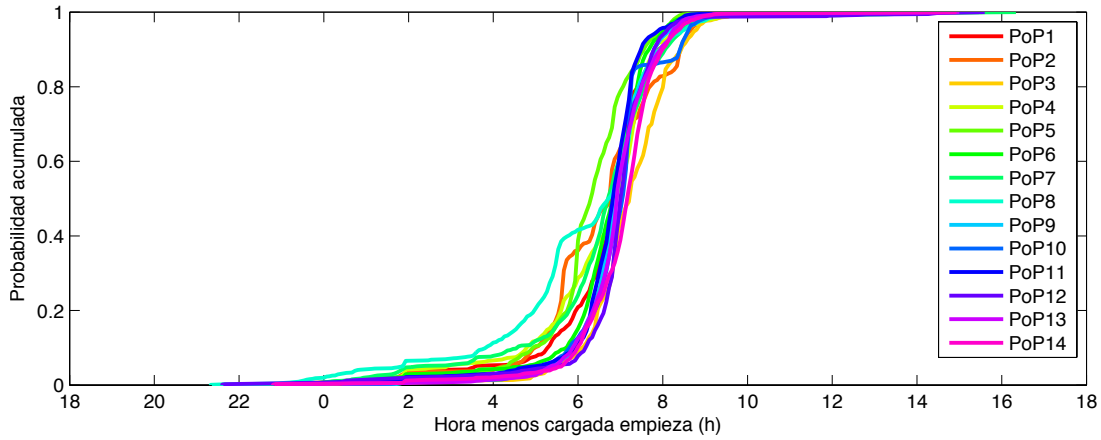

Figura 6-1: Superposición de FDAs de todos los PoPs

La superposición de las FDAs hace palpable la semejanza entre todos los exportadores, que concentran sus horas de menor utilización entre las 5 y las 9 horas aproximadamente y apenas cuentan con momentos valle fuera de ese lapso de tiempo.

Otra forma de ver este efecto es según estimaciones de la FDP de cada exportador, expuestas mediante un suavizado de tipo gaussiano en la Figura 6-2.

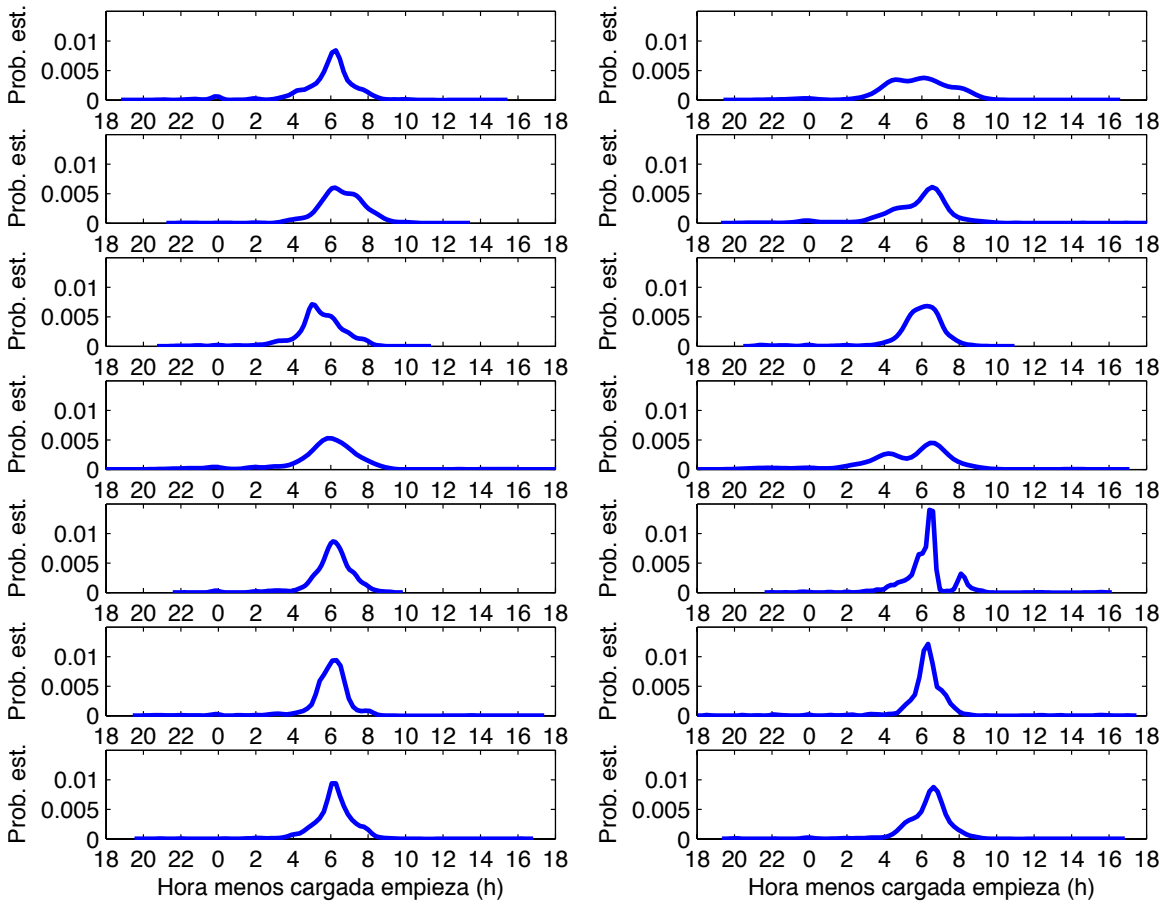

Figura 6-2: FDPs de todos los PoPs

De nuevo, se observan medias y modas claramente similares, aunque esta representación permite detectar mejor las pequeñas singularidades presentes en cada uno de los PoPs. Así, se intuyen ciertos rasgos característicos de cada subred que probablemente impiden rechazar la no relevancia del factor exportador.

6.2 Análisis ANOVA

6.2.1 Introducción

La primera herramienta empleada en la predicción que se propone es ANOVA o análisis de varianza, de tipo univariante, es decir, con una única variable dependiente u observable. ANOVA es una técnica estadística para analizar y explicar medidas a partir de varios efectos simultáneos para decidir cuáles son significativos y cuantificar su impacto [43], adaptada a una gran variedad de experimentos.

Su empleo conlleva varios supuestos [44]: independencia de las observaciones, distribución normal de las mismas y homocedasticidad, o lo que es lo mismo, igualdad de varianzas. Adicionalmente, se requiere que el experimento no tenga tendencia con el tiempo, cuestión discutida con anterioridad en la Subsección 4.4, que las observaciones cuenten con una población -número de muestras- equitativa o muy próxima a ello, aspecto tenido en cuenta en 4.3.2.

En cuanto a la homogeneidad de las varianzas, su rigurosidad se ve relajada en algunas publicaciones [45], que sugieren que una heterocedasticidad moderada con un número grande de muestras como se da en nuestro problema puede relajar este supuesto a la hora de establecer el análisis, si bien los resultados se ven levemente afectados por este hecho.

Para el caso que nos ocupa, se opta por un modelo de efectos fijos, en el que la variable observable, la hora menos cargada, se expresa en función de variables explicativas, tomadas como no aleatorias. Este tipo de análisis atañe a situaciones en las que los factores afectan únicamente a la media.

6.2.1 Descripción

ANOVA plantea la hipótesis nula [46] de que observaciones en principio clasificables en grupos provienen de la misma población, amparada en la igualdad de medias (μ). Esta hipótesis se formula de la siguiente manera:

$y_{ij} = \mu + u_{ij}$, siendo y_{ij} los valores observados y u_{ij} el error experimental.

Frente a la cual se enuncia a una hipótesis alternativa de medias diferentes:

$$y_{ij} = \mu_i + u_{ij}$$

La variabilidad, definida como la desviación entre los datos observados y la media, puede descomponerse en dos fuentes de variación, entre grupos (VG) -explicada- y residual (VR) -no explicada-. De esta descomposición resulta el coeficiente de determinación, R^2, una medida relativa de la variabilidad que podemos definir como:

$$R^2 = \frac{VG}{VR}$$

Este valor, entre 0 y 1, nos permite calibrar el grado de explicación del fenómeno. Cuanto mayor sea el número, más exacto y fidedigno será el análisis ANOVA.

El test calcula si un factor es significativo o no según la distribución F de Fisher-Snedecor, el cociente entre la varianza de VG y la varianza VR. Si ambas varianzas son similares -cociente próximo a 1-, se entenderá que ese factor no es significativo; en caso contrario, se deducirá que el factor es significativo.

El test presenta su veredicto de aprobación o no de la hipótesis como un valor denominado *Sig.*, de acuerdo a un nivel de significación α estipulado; típicamente y en este trabajo, igual a 0’05. Un parámetro Sig. inferior a α indicará un factor significativo; no significativo en caso de ser superior a α.

6.3 Predicción para un único PoP

El pronóstico de la hora valle en una subred individual se afronta de dos maneras. La primera se apoya en ANOVA, que según lo dispuesto en la subsección anterior es presumible que produzca una predicción con ciertas inexactitudes. Como segunda opción se expone la predicción a partir de los intervalos de confianza de una distribución normal.

6.3.1 Estimación mediante ANOVA

La variable dependiente que evalúa el test es el *minuto hora menos cargada*, el momento en el que se inicia la hora valle. Asimismo, los factores fijos escogidos que darán lugar a los diferentes grupos son exportador, día de la semana, mes y la condición del día, entre laborable o festivo. Factores inicialmente planteados, como año o ancho de banda durante la hora valle, se descartan, respectivamente, por la condición de estacionariedad y la no correlación del ancho de banda con la hora valle descritos en la Sección 4.

Como covariables, o variables externas que se intuye que pueden ayudar a la explicación del fenómeno -y de hecho se mejora su incidencia positiva en ello- y de paso probar su significación o no en el problema: IPs activas durante la hora menos cargada, IPs activas a lo largo del día y ancho de banda medio durante todo el día.

En la Figura 6-3 se muestra el resultado del análisis ANOVA.

Variable dependiente: Minuto_hora_menos_cargada

Origen	Tipo III de suma de cuadrados	gl	Cuadrático promedio	F	Sig.
Modelo corregido	38041739[a]	1550	24543,057	2,468	,000
Interceptación	89873150,5	1	89873150,5	9039,151	,000
IPs_activas_hora_menos_cargada	112069,010	1	112069,010	11,272	,001
IPs_activas_dia	293251,632	1	293251,632	29,494	,000
Mbits_medios_dia	520857,277	1	520857,277	52,386	,000
Exporter	4738718,52	13	364516,809	36,662	,000
MonthL	966077,242	11	87825,204	8,833	,000
Weekday	1362855,34	6	227142,556	22,845	,000
LF	96499,248	1	96499,248	9,706	,002
Exporter * MonthL	1943419,98	143	13590,349	1,367	,003
Exporter * Weekday	1678592,09	78	21520,411	2,164	,000
Exporter * LF	208638,042	13	16049,080	1,614	,073
MonthL * Weekday	1371506,41	66	20780,400	2,090	,000
MonthL * LF	143814,571	7	20544,939	2,066	,044
Weekday * LF	109638,937	4	27409,734	2,757	,026
Exporter * MonthL * Weekday	8874812,01	843	10527,654	1,059	,123
Exporter * MonthL * LF	776236,595	76	10213,639	1,027	,413
Exporter * Weekday * LF	547683,920	52	10532,383	1,059	,359
MonthL * Weekday * LF	207451,065	22	9429,594	,948	,529
Exporter * MonthL * Weekday * LF	2195548,07	197	11144,914	1,121	,120
Error	119779158	12047	9942,654		
Total	1,840E+9	13598			
Total corregido	157820897	13597			

a. R al cuadrado = ,241 (R al cuadrado ajustada = ,143)

Figura 6-3: Resultado del análisis ANOVA

Como vemos, se obtiene un valor R^2 bajo -0‘241-, que no permite caracterizar plenamente el fenómeno y, por tanto, deparará resultados y predicciones no óptimos. Este valor se explica de muchas formas: la no homocedasticidad exacta en nuestro problema, el elevado número de muestras totales, la ineficacia de los factores para describir el fenómeno... y dirige a la propuesta de un sistema de predicción distinto en la Subsección 6.3.3.

Pese a ello, los datos se toman como suficientemente válidos para extraer conclusiones. ANOVA corrobora la significación de los factores externos introducidos, el número de IPs activas durante la hora menos cargada, el número de IPs activas a lo largo del día y ancho de banda medio durante todo el día. De esta manera, podemos concluir que estos tres elementos no afectan al momento en el que tiene lugar la hora valle.

Asimismo, el análisis revela que los cuatro factores fijos han de considerarse significativos. Creemos que esto es debido a la gran cantidad de datos de los que consta nuestro experimento, que supone que pequeñas diferencias en las medias de las poblaciones sean evaluadas como significativas. La principal consecuencia radica en la contemplación de las subredes de forma aislada, ya que cada una de ellas debe ser estudiada por separado.

Con el objetivo de intentar proporcionar un estudio de la varianza más certero, se medita crear nuevos grupos de exportadores, días de la semana y meses a través de la prueba de Bonferroni. Este test permite comparar de modo múltiple, unas frente a otras, cada una de las clases que componen los factores fijos para detectar grupos, que pueden servir para aumentar el valor R^2. Así, tomando los resultados de la prueba se dividen los meses y los exportadores en tres grupos respectivos y los días de la semana en dos grupos. Pero esta estrategia no resulta satisfactoria: no sólo no logramos una mejor explicación del fenómeno, si no que R^2 se reduce. Consecuentemente, esta opción es descartada.

La predicción se efectúa de forma práctica mediante la estimación de los parámetros, un método incluido en ANOVA que permite dar un resultado para la predicción a partir de unos coeficientes calculados para cada uno de los factores. Siendo la media μ, y asignando α al factor exportador, β al factor mes, δ al factor día de la semana y φ al factor laboral/festivo, la predicción ‘y’ se obtiene como:

$$y = \mu + \alpha + \beta + \delta + \varphi + \alpha\beta + \alpha\delta + \alpha\varphi + \beta\delta + \beta\varphi + \delta\varphi + \alpha\beta\delta + \alpha\beta\varphi + \alpha\delta\varphi + \beta\delta\varphi + \alpha\beta\delta\varphi$$

Cada uno de los coeficientes presenta un error estándar que, por las excepcionalidades comentadas con anterioridad en esta sección, son relativamente relevantes, aunque por simplicidad y por tratarse de una estimación teórica, no se tendrán en cuenta.

En la Tabla 6-1 se ejemplifican cuatro predicciones de la hora valle para un único PoP siguiendo el procedimiento estimación de parámetros de ANOVA.

Caso	Exportador	Día de la semana	Mes	L/F	Predicción hora valle
1	PoP3	Lunes	Enero	L	06:09h
2	PoP6	Jueves	Octubre	L	05:56h
3	PoP11	Viernes	Mayo	L	05:34h
4	PoP14	Domingo	Julio	F	06:53h

Tabla 6-1: Ejemplos de predicción mediante ANOVA

Los resultados de la predicción se encuentran en consonancia con la visión general de las FDAs. En los cuatro casos se estima que el momento en el que empieza la hora de menor utilización de las subredes está en torno a las 6 horas, que coincide con los puntos de la mayor pendiente de las distribuciones acumuladas.

La circunstancia de que la hora valle se localice en la zona cercana a la moda constituye un punto a favor de la exactitud de la predicción, puesto que en una distribución normal como la de dicha variable, acercarse a la moda implica aumentar progresivamente la precisión del resultado.

6.3.2 Discusión del efecto de los parámetros

La predicción expuesta en la Subsección 6.3.1 utiliza un ANOVA de tipo factorial completo. Esta pauta para el análisis se apoya, con la finalidad de aumentar la explicación del fenómeno, en el cómputo de la significación de la multiplicación de todos los elementos que constituyen los factores.

En contraposición, contabilizar la contribución de cada una de las combinaciones entre parámetros hace que se difumine el efecto que tienen individualmente sobre el pronóstico de la hora valle. Este aspecto se considera de gran valor para deducir algún tipo de comportamiento coincidente entre distintos grupos.

Por este motivo, se procede a evaluar nuevamente nuestro problema mediante ANOVA, pero esta vez centrado en un análisis por efectos principales. No se puede obviar que en este caso el valor de R^2 mengua, por lo cual la precisión del modelo es algo más baja que cuando se trabaja con un análisis factorial completo. Esto no es impedimento para que las apreciaciones más evidentes puedan ser consideradas como razonables.

El impacto de los parámetros más considerable es el de los fines de semana. ANOVA estima que el factor fin de semana provoca que la hora menos cargada tenga lugar aproximadamente 55 minutos después que un día cualquiera entre semana. Aunque en una red académica sobre la que nos basamos sea complejo explicar las causas tras este hecho, una posibilidad recurrente es que los usuarios -buena parte de ellos, probablemente, los que residen en las universidades- se acuestan y despiertan por lo general más tarde, desplazando el momento valle unos minutos más adelante en el día.

En cuanto a los meses, se estima que la hora valle ocurre unos quince o veinte minutos antes en los meses de vacaciones de verano -julio y agosto- que en los meses centrales del curso académico. Como hipótesis, planteamos el menor acceso a las primeras horas del día como explicación a este hecho. Por el contrario, no se distinguen conductas parejas entre las distintas subredes, cuyas horas valle se reparten en un rango más o menos amplio, ni tampoco para la distinción entre laboral y festivo, con una diferencia que apenas roza los cinco minutos.

6.3.3 Estimación mediante intervalo de confianza

Una práctica común cuando se ha probado la normalidad de la población de un experimento es una estimación por intervalos, menos determinista pero a la vez más

rigurosa, debido a que permite precisar la incertidumbre existente en la estimación [47]. Esta forma de enfocar el problema tendrá también relevancia cuando afrontemos, en la Subsección 6.4, el problema de transmisiones entre PoPs que requieran saltos intermedios para llegar de origen a destino.

Este método posibilita establecer un intervalo con un nivel de confianza de 1 - α, para un grado de significación, en nuestro problema definido como α = 0'05. Con él, podemos fijar unos valores mínimo y máximo entre los que se encontrará la predicción, un intervalo que será acertado con una probabilidad del 95%.

Dada la normalidad de nuestra variable hora valle y conocida su media muestral $\bar{x}$ y su desviación típica σ a través de las muestras empíricas tomadas en los seis años que abarca nuestro estudio, el intervalo al 95% para la media se puede definir como [35]:

$$\bar{x} - 1'96\sigma \leq \mu \leq \bar{x} + 1'96\sigma$$

Para este tipo de tareas con exigencias bajas en cuanto a *timing*, este margen óptimo para el 95% de las veces se ajusta adecuadamente y permite una predicción sencilla cuando tratamos de estimar el momento valle.

Dicha expresión se evalúa para nuestro problema. La media muestral y la varianza se calculan a partir de los días existentes que comparten los factores de exportador, día de la semana, mes y laboral/festivo que buscamos. Para posibilitar la comparación con el método ANOVA, se escogen los mismos cuatro casos que en la Subsección 6.3.1.

La Tabla 6-2 muestra los resultados de la predicción para los cuatro ejemplos.

Caso	Exportador	Día de la semana	Mes	L/F	Media	Intervalo
1	PoP3	Lunes	Enero	L	06:15h	04:05-08.25h
2	PoP6	Jueves	Octubre	L	06:02h	04:51-07:11h
3	PoP11	Viernes	Mayo	L	05:43h	04:11-07.15h
4	PoP14	Domingo	Julio	F	06:53h	05:18-08:30h

Tabla 6-2: Ejemplos de predicción mediante intervalo de confianza

Apreciamos que la predicción es concordante con la decretada por ANOVA. El valor de la media es muy similar, e incluso idéntica para el caso 4, y los intervalos que se establecen tienen un rango de aproximadamente tres o cuatro horas. El intervalo es el mecanismo mediante el cual podemos, de algún modo, estimar la fiabilidad de la media, de manera que un rango más pequeño nos permite aseverar con mayor rotundidad la media como el momento idóneo. También permite ofrecer cierta libertad al gestor de la red en cuanto a la planificación las transferencias de tipo *bulk*, objeto de este estudio.

No obstante, debemos tener presente la trascendencia del número de muestras a partir de las cuales se calculan las medias empíricas y las varianzas. En los cuatro ejemplos presentados previamente contamos con entre cinco y diez muestras por caso, lo cual implica que una única muestra distorsionada por cualquier eventualidad en un día puede afectar notablemente al tamaño del intervalo, aunque su validez se mantendría intacta.

6.4 Predicción para una ruta por varios PoPs

6.4.1 Motivación

Tras exponer dos procedimientos recurrentes para la predicción referida a un único exportador, se entiende como inmediato dar un paso más en el análisis y disponer de diversas formas de afrontar un problema sustancialmente más complejo: la predicción de la hora valle para un enlace o *path* entre dos o más subredes.

El enlace se construye a partir de varios PoPs a modo de prototipo teórico y conceptual, sin reparar en la existencia o no de un enlace físico entre los PoPs en RedIRIS, siguiendo la tónica de privacidad de los datos de las subredes que rige este estudio. Esto no es óbice para que las diferentes estrategias planteadas sean plenamente extrapolables a una conexión real.

El análisis asume que la incorrelación entre la hora valle y el ancho de banda -Subsección 4.5- supone que, sobre el papel, no se pueda decretar ninguna clase de prioridad intrínseca a las subredes. De valorar algún tipo de preponderancia entre unos PoPs y otros, esta se fijará arbitrariamente.

6.4.2 Modelo NetStitcher

La escasa restricción en cuanto a horarios concretos en tareas de tipo *bulk* da origen a una de las formas de atajar el análisis más simples para un enlace, consistente en que cada uno de los exportadores efectúe la conexión en su momento más apropiado. Esta opción ha sido contemplada en estudios similares para maximizar el volumen de labores de *backup* en *data centers* [18].

En esencia, el funcionamiento es el que sigue: el PoP situado en el origen del *path* inicia la transmisión en su hora valle idónea, cuando llega al siguiente PoP éste aguarda al momento de su hora valle para hacer el envío -esperando al día posterior si es necesario- y así sucesivamente, hasta que finalmente se alcanza el PoP destino.

El hecho de que la conexión se quede *congelada* durante casi un día, si el momento valle de un *router* que va a emitir ha ocurrido hace escasos minutos, conlleva algunos inconvenientes, sobre todo pensando en enlaces con muchos puntos intermedios, cuyas tareas *bulk* podrían alargarse un tiempo considerable, más del deseable. Por este motivo, su utilización puede verse limitada a enlaces con no muchos saltos.

Para ilustrar esta metodología, se especifican las predicciones para dos días consecutivos del mismo mes de varios PoPs, reflejados en la Tabla 6-3.

Exportador	Día de la semana	Mes	L/F	Predicción hora valle
PoP1	Lunes	Noviembre	L	05:23h
PoP2	Lunes	Noviembre	L	06:08h

PoP3	Lunes	Noviembre	L	06:13h
PoP1	Martes	Noviembre	L	05:42h
PoP2	Martes	Noviembre	L	06:09h
PoP3	Martes	Noviembre	L	06:24h
PoP1	Miércoles	Noviembre	L	05:34h

Tabla 6-3: Predicciones de horas valle de PoP1, PoP2 y PoP3 (ANOVA)

A partir de estos datos se proponen varios ejemplos de *paths*. Despreciando los posibles retardos en el enlace físico entre PoPs, en la Tabla 6-4 se calcula el tiempo que transcurre desde que un paquete de una conexión de tipo *bulk* sale el lunes de un PoP origen, pasa por un PoP intermedio y finalmente llega a un PoP destino.

Ruta	PoPs	Retardo origen-destino
1	PoP1-PoP2-PoP3	50m
2	PoP1-PoP3-PoP2	24h46m
3	PoP2-PoP1-PoP3	24h16m
4	PoP2-PoP3-PoP1	23h34m
5	PoP3-PoP1-PoP2	23h56m
6	PoP3-PoP2-PoP1	47h21m

Tabla 6-4: Ejemplos de retardos en *paths* siguiendo el método NetStitcher

En estos casos -que engloban todas las combinaciones posibles para estas subredes PoP1, PoP2 y PoP3 específicas- de *paths*, comprobamos que el 16'7% de las veces el paquete llega en menos de una hora, el 66'7% de las veces alrededor de un día y el 16'7% restante en torno a dos días. Las proporciones para este ejemplo nos dan una idea de cómo se pueden distribuir los retardos para tres exportadores cualesquiera.

6.4.3 Compendio de distribuciones normales

La estimación mediante intervalos de confianza al 95% detallada en 6.3.3 y las propiedades de una distribución normal nos permiten establecer la predicción como un promedio de la suma de las tres variables aleatorias correspondientes a las horas valle de cada PoP por separado. Así, podemos predecir una hora valle *global,* que tiene en consideración las medias y los intervalos de confianza individuales.

Siendo X, Y y Z tres variables aleatorias, normales e independientes, con medias μ_x, μ_y y μ_z y varianzas ${\sigma_x}^2, {\sigma_y}^2$ y ${\sigma_z}^2$ respectivamente, sumamos X, Y y Z para obtener S = X + Y + Z.

El resultado S es también una variable aleatoria normal y puede escribirse como S ~ $N(\mu_x + \mu_y + \mu_z, {\sigma_x}^2 + {\sigma_y}^2 + {\sigma_z}^2)$ [48]. Si la multiplicamos por un número a real: $a*N(\mu_x + \mu_y + \mu_z, {\sigma_x}^2 + {\sigma_y}^2 + {\sigma_z}^2) = N(a * (\mu_x + \mu_y + \mu_z), a^2({\sigma_x}^2 + {\sigma_y}^2 + {\sigma_z}^2))$. Definiendo el número 'a' como 0'33 obtenemos un promedio de la distribución.

En un primer acercamiento, esta aproximación se evalúa para tres rutas balanceadas diferentes, que a su vez atraviesan tres PoPs cada una. Las predicciones para cada PoP en particular se muestran en la Tabla 6-5.

Exportador	Día de la semana	Mes	L/F	Media	Intervalo
PoP1	Martes	Noviembre	L	05:45h	04:54-06:37h
PoP4	Martes	Noviembre	L	05:40h	03:38-07:42h
PoP9	Martes	Noviembre	L	05:50h	04:34-07:06h
PoP3	Sábado	Febrero	F	06:49h	03:55-09:44h
PoP6	Sábado	Febrero	F	06:21h	04:22-08:19h
PoP13	Sábado	Febrero	F	07:14h	05:29-08:58h
PoP2	Lunes	Diciembre	L	06:12h	04:36-07:49h
PoP5	Lunes	Diciembre	L	05:20h	03:55-06:45h
PoP8	Lunes	Diciembre	L	04:21h	00:54-07:48h

Tabla 6-5: Ejemplos de predicción mediante intervalo de confianza

Destaca la gran variabilidad en las horas valle de PoP8, que con toda probabilidad supondrán un efecto claro en su ruta. La Tabla 6-6 muestra la predicción para las rutas.

Ruta	PoPs	Media	Intervalo
1	PoP1-PoP4-PoP9	05:42h	04:51-06:32h
2	PoP3-PoP6-PoP13	06:44h	05:26-08:01h
3	PoP2-PoP5-PoP8	05:15h	03:54-06:35h

Tabla 6-6: Ejemplos de predicción para un path *equilibrado* mediante intervalo de confianza

Nótese que, al contrario que en el modelo NetStitcher en 6.4.2, en este punto es indiferente el sentido de la ruta, ya que la predicción del momento valle tiene en cuenta por igual a los tres PoPs.

Un segundo prototipo de predicción mediante intervalo de confianza para varios PoP reside en determinar una predicción análoga, pero simulando la intervención de un gestor de red. Esta figura asigna pesos que otorgan mayor importancia en el cálculo de la media y en el intervalo de la hora valle conjunta, ya sea por motivos de utilización, coste o cualquier otro factor arbitrario.

En la Tabla 6-7 se repite la ruta anterior, pero con los resultados de fijar pesos. Los PoPs situados más a la izquierda en la fila tienen un factor 6 -son los de más peso-, los situados en medio un factor 3 y los situados a la derecha un factor 1 -los de menos peso-.

Ruta	PoPs	Media	Intervalo
1	PoP1-PoP4-PoP9	05:44h	04:56-06:33h
2	PoP3-PoP6-PoP13	06:43h	04:52-08:34h
3	PoP2-PoP5-PoP8	05:45h	04:39-06:52h

Tabla 6-7: Ejemplos de predicción para un *path* con pesos mediante intervalo de confianza

El efecto de los pesos se hace notar, sobre todo en la tercera ruta, en la que PoP8 -el que depara un intervalo para la hora valle más dispar- pasa a tener mucha menor trascendencia que en el caso balanceado. Las otras dos experimentan cambios leves, pues sus distribuciones son mucho más similares.

6.4.4 Modelo de *metanodos*

La última propuesta de predicción en este trabajo es un modelo, de nuevo, basado en el intervalo de confianza al 95%, pero esta vez viendo la ruta entre tres PoPs con un único *metanodo*. Esta forma de afrontar el pronóstico supone calcular el momento valle a partir de los tres PoPs a la vez, o lo que es lo mismo, teniendo en consideración el ancho de banda de cada uno de ellos para dictaminar el instante en el que se inicia la hora valle.

Una de las dificultades que entraña esta conceptuación es que el número de muestras empíricas se ve reducida, como consecuencia de que si se detecta algún problema -tanto por falta de días como por tolerancia a errores- en alguno de los días o en su siguiente -idea expuesta en la Subsección 4.1.1- en cualquiera de los tres PoP, todo ese día es eliminado del análisis. A su favor, se trata de un modelo que evalúa un aspecto de interés en el análisis como es la utilización de cada subred, pues estas pueden tener anchos de banda diferentes.

En la Tabla 6-8 se vuelve a tomar los datos de la Tabla 6-5 para, como en la Tabla 6-6 y Tabla 6-7, decretar la predicción para las tres rutas.

Ruta	PoPs	Media	Intervalo
1	PoP1-PoP4-PoP9	05:40h	03:38-07:42h
2	PoP3-PoP6-PoP13	06:59h	04:42-09:15h
3	PoP2-PoP5-PoP8	04:43h	02:31-06:54h

Tabla 6-8: Ejemplos de predicción para un *metanodo* mediante intervalo de confianza

Apreciamos que, respecto a la propuesta de la subsección anterior, la media prácticamente no se altera para las dos primeras rutas, aunque sí en la tercera, debido a en que en este caso disminuye notablemente el número de muestras. Esto también se percibe los intervalos de los tres casos, que aumentan y, con ellos, la incertidumbre del resultado.

7 Conclusiones y trabajo futuro

7.1 Conclusiones

Este trabajo de fin de grado ha acreditado la importancia de realizar un estudio acerca del momento idóneo para la realización de transferencias de tipo *bulk*, amparado por otras publicaciones previas que han deliberado acerca de su problemática, especialmente referida a centros de datos y modelos de pago como el modelo percentil 95.

Tras una reseña de la tecnología y de la infraestructura que ha permitido el análisis, hemos completado una supervisión de los datos disponibles, repartida en dos fases. En la primera, se han descartado tres PoPs por su escasez de medidas; en la segunda; otros tres exportadores se han dejado a un lado por las dudas acerca de la integridad de las muestras.

Sin embargo, algunas singularidades han sido aceptadas, con el objetivo de obrar un análisis amplio y genérico. El examen de los errores revela que su incidencia en la distribución de la hora valle es, por lo general, baja, aunque se han encontrado peculiaridades.

La estacionariedad del problema ha sido expuesta de manera visual, comprobando la ausencia de tendencias evidentes. Además, hemos detectado cambios con el paso del tiempo en el ancho de banda, pero se ha comprobado su incorrelación con el objeto del trabajo, como también la del momento de mayor utilización.

La caracterización más apropiada ha resultado ser de tipo normal o gaussiana, justificada, tras la transformación de los datos, por un test de banda de ajuste. Este modelado nos ha permitido acometer la predicción de la hora valle, abordada por un doble camino: un análisis ANOVA y un procedimiento mediante intervalos de confianza. A través de la visualización de todos los exportadores en conjunto, hemos localizado que la mayoría de las horas valle tienen lugar entre las 5 y las 9 horas y la analogía entre subredes.

ANOVA nos ha permitido considerar tanto el día de la semana, el mes, el exportador y la condición de laboral o festivo como factores significativos, probablemente como consecuencia del tamaño de la población de nuestro experimento. Asimismo, hemos podido constatar las dificultades para encontrar invariantes explorando la existencia de algún parámetro con impacto notorio en el resultado. En este sentido, se ha deducido principalmente que la hora valle ocurre casi una hora después en fines de semana frente al resto de los días.

Como alternativa y dada la normalidad de nuestra variable, hemos introducido el concepto de intervalo de confianza y ofrecido predicciones según la media de las muestras empíricas, cuya variabilidad es cuantificada mediante el intervalo, calculado a partir de la varianza.

El último planteamiento desarrollado ha sido la predicción para una ruta entre distintos PoPs, de acuerdo a varios métodos. El primero de ellos ha seguido, mediante ANOVA, el proceder de NetStitcher, apoyado en las exigencias bajas de *timing* en las tareas *bulk*. Otra

aproximación ha consistido en calcular un promedio entre distribuciones, modificado para simular la intervención de un gestor de red en la conexión. Como colofón, hemos formulado un modelo de *metanodos*, en los que el ancho de banda de cada subred intensifica su papel en la predicción.

Las comparaciones que se han podido establecer entre las distintas predicciones han mostrado resultados esencialmente equitativos, lo que nos lleva a concluir que hemos alcanzado el objetivo de ofrecer una predicción ajustada para la hora valle.

7.2 Trabajo futuro

La propuesta de la hora valle detallada en este trabajo abre la puerta a publicaciones futuras que, partiendo de la línea trazada en este análisis, enriquezca la caracterización del fenómeno y comparta las bondades de la propuesta de la hora valle como elemento clave en el problema del *timing* de actividades tipo *bulk*.

Para ello, se considera primordial ser capaces de caracterizar más a fondo las subredes. Más concretamente, sería deseable encontrar más factores que permitieran modelar la variable hora valle con menor incertidumbre.

Por otro lado, sería de interés indagar en las peculiaridades que hemos encontrado en algunas subredes durante nuestra caracterización en días concretos y, a veces, durante periodos significativos. En particular, nos referimos a valores anómalos detectados tras una simple inspección visual; por ejemplo, incrementos del uso de las redes a altas horas de la madrugada. Especulamos que se trata de propias tareas *bulk* de las redes o conexiones muy puntuales de centros de investigación transfiriendo datos de un experimento. Se trata de conjeturas a las que podríamos dar un respuesta firme analizando el tráfico a nivel de paquete.

Finalmente, nos planteamos como trabajo futuro llegar a implementar nuestra propuesta como una aplicación real o *plug-in*, y facilitarla libremente para su uso por gestores de red. Idealmente, el estudio de su uso en un despliegue grande sería muy enriquecedor.

Referencias

[1] N. Laoutaris, G. Smaragdakis, P. Rodriguez and R. Sundaram, 'Delay-Tolerant Bulk Data Transfers on the Internet'. IEEE/ACM Transactions on Networking (TON), 2013, vol. 21, no 6, p. 1852-1865.

[2] Forrester Research, 'The future of data center wide-area networking'. http://www.forrester.com.

[3] S. Floyd and V. Paxson, 'Difficulties in Simulating the Internet', IEEE/ACM Transactions on Networking (TON), 2001, vol. 9, no 4, p. 392-403.

[4] J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara and S. López-Buedo, 'Characterization of the busy-hour traffic of IP networks based on their intrinsic features'. Computer Networks, 2011, vol. 55, no 9, p. 2111-2125.

[5] V. Gupta, 'What is network planning?'. Communications Magazine, IEEE, 1985, vol. 23, no. 10, p. 10-16.

[6] X. Dimitropoulos, P. Hurley, A. Kind and M. P. Stoecklin, 'On the 95-Percentile Billing Method'. Passive and Active Network Measurement. Springer Berlin Heidelberg, 2009. p. 207-216.

[7] 'Introduction to Cisco IOS NetFlow', http://www.cisco.com.

[8] D. López, J. E. López-de-Vergara, L. Bellido and D. Fernández, 'Monitoring an academic network with Netflow'. Proceedings of EUNICE. 2004.

[9] A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras, and B. Stiller, 'An overview of IP flow-based intrusion detection'. Communications Surveys & Tutorials, IEEE, 2010, vol. 12, no 3, p. 343-356.

[10] V. Carela-Espanol, P. Barlet-Ros and J. Solé-Pareta, 'Traffic classification with sampled netflow'. traffic, 2009, vol. 33, p. 34-34.

[11] R. Hofstede, P. Celeda, B. Trammell, I. Drago, R. Sadre, A. Sperotto and A. Pras, 'Flow monitoring explained: from packet capture to data analysis with netFlow and IPFIX'. IEEE Communications Surveys & Tutorials, 2014, vol. 16, no. 4, pp. 2037–2064.

[12] W. Van Wanrooij and A. Pras, 'Data on Retention'. Ambient Networks. Springer Berlin Heidelberg, 2005. p. 60-71.

[13] C. Estan and G. Varghese, 'New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice'. ACM Transactions on Computer Systems (TOCS), 2003, vol. 21, no 3, p. 270-313.

[14] Chelo Malagón, ‘NetFlow y su aplicación en seguridad’. Enfoques, http://www.rediris.es/difusion/publicaciones/boletin/87/enfoque1.pdf

[15] L. Deri and N. SpA, 'nProbe: an open source netflow probe for gigabit networks'. Proceedings of Terena TNC, 2003.

[16] http://www.splintered.net/sw/flow-tools

[17] V. Moreno, P. M. Santiago del Río, J. Ramos, D. Muelas, J. L. García-Dorado, F. J. Gómez-Arribas and J. Aracil, 'Multi-granular, Multi-purpose and Multi-Gb/s Monitoring on Off-the-shelf Systems'. International Journal of Network Management, 2014, vol. 24, no 4, p. 221-234.

[18] N. Laoutaris, M. Sirivianos, X. Yang, and P. Rodriguez, 'Inter-datacenter bulk transfers with netstitcher'. ACM SIGCOMM Computer Communication Review. ACM, 2011. p. 74-85.

[19] L. Gyarmati, R. Stanojevic, M. Sirivianos, and N. Laoutaris, 'Sharing the cost of backbone networks: cui bono?'. Proceedings of the 2012 ACM conference on Internet measurement conference. ACM, 2012. p. 509-522.

[20] Y. Feng, B. Li and B. Li, 'Postcard: Minimizing costs on inter-datacenter traffic with store-and-forward'. Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on. IEEE, 2012. p. 43-50.

[21] T. Nandagopal and K. P. Puttaswamy, 'Lowering inter-datacenter bandwidth costs via bulk data scheduling'. Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on. IEEE, 2012. p. 244-251.

[22] J. L. García-Dorado, J. A. Hernández, J. Aracil and J. E. López de Vergara, ‘On the Duration and Spatial Characteristics of Internet Traffic Measurement Experiments’. Communications Magazine, IEEE, 2008, vol. 46, no 11, p. 148-155.

[23] K. Papagiannaki, N. Taft, Z. L. Zhang and C. Diot, 'Long-term forecasting of internet backbone traffic: Observations and initial models'. INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies. IEEE, 2003. p. 1178-1188.

[24] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry and K. Cho, 'Seven years and one day: Sketching the evolution of internet traffic'. INFOCOM 2009, IEEE. IEEE, 2009. p. 711-719.

[25] F. Mata, J. L. García−Dorado, J. Aracil and J. E. López de Vergara, ‘Factor analysis of Internet traffic destinations from similar source networks’. Internet Research, 2012, vol. 22, no 1, p. 29-56.

[26] http://www.hpcn.es/project/anfora-plan-nacional-de-id

[27] http://www.rediris.es/rediris/

[28] http://www.rediris.es/rediris/historia/programa-iris.pdf

[29] http://www.redirisnova.es/antecedentes.html

[30] http://www.redirisnova.es/mm/presentacion-RedIRIS-NOVA.pdf

[31] Peyton Z. Peebles, 'Principios de probabilidad, variables aleatorias y señales aleatorias'. McGraw-Hill, 4ª ed.

[32] Curso Combinado de Predicción y Simulación, 'Unidad 6: Modelos Econométricos Uniecuacionales'. http://www.uam.es/docencia/predysim/

[33] M. S. Kim, Y. J. Won, J. W. Hong, 'Characteristic analysis of internet traffic from the perspective of flows'. Computer Communications, 2006, vol. 29, no 10, p. 1639-1652.

[34] D. Peña Sánchez de Rivera, 'Fundamentos de Estadística'. Alianza Editorial, 2001.

[35] George C. Canavos, 'Probabilidad y Estadística: Aplicaciones y Métodos'. McGraw-Hill, 1ª ed, p. 137.

[36] http://es.wikipedia.org/wiki/Distribución_normal. Licencia CC BY-SA 3.0.

[37] http://es.wikipedia.org/wiki/Distribución_uniforme_continua

[38] R. B. D'Agostino and M. A. Stephens, 'Goodness-of-Fit Techniques'. Marcel Dekker, Inc., 1986.

[39] H.W. Lilliefors, 'On the Kolmogorov-Smirnov test for normality with mean and variance unknown'. Journal of the American Statistical Association, 1967, vol. 62, no 318, p. 399-402.

[40] J. Kilpi and I. Norros, 'Testing the Gaussian approximation of aggregate traffic', Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment. ACM, 2002. p. 49-61.

[41] R. van de Meent, M. Mandjes and A. Pras, 'Gaussian traffic everywhere?'. Communications, 2006. ICC'06. IEEE International Conference on. IEEE, 2006. p. 573-578.

[42] D. Peña Sánchez de Rivera, 'Regresión y diseño de experimentos'. Alianza Editorial, 2002.

[43] Henry Scheffé, 'The analysis of variance'. John Wiley & Sons, 1959.

[44] Olive Jean Dunn, 'Applied statistics analysis of variance and regression'. Wiley, 1974.

[45] G.V. Glass, P. Peckham and J.R. Sanders, 'Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance'. Review of educational research, 1972, p. 237-288.

[46] D. Peña Sánchez de Rivera, ‘Estadística: modelos y métodos. 2. Modelos lineales y series temporales’. Alianza Editorial, 2ª ed., 1989.

[47] D. Peña Sánchez de Rivera, ‘Estadística: modelos y métodos. 1. Fundamentos’, Alianza Editorial. 2ª ed., 1986.

[48] E. W. Weisstein, ‘Normal Sum Distribution’. From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/NormalSumDistribution.html

Glosario de abreviaturas

ANFORA	Análisis forense, longitudinal y ciego de tráfico de Internet
ANOVA	Analysis of variance
API	Application Programming Interface
DIOR	Dimensionado de redes IP y redes ópticas
FE	Frontera de error
FDA	Función de distribución acumulada
FDP	Función de densidad de probabilidad
GMT	Greenwich Mean Time
IP	Internet Protocol
IPFIX	Internet Protocol Flow Information Export
IRIS	Interconexión de Recursos Informáticos
ISP	Internet Service Provider
MAWI	Measurement and Analysis on the Wide Internet
PoP	Point of presence
RAM	Random-access memory
SNMP	Simple Network Management Protocol
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VG	Variación entre grupos
VR	Variación residual

Anexos

A Tabla de archivos disponibles

Año-mes/ Exporter	Badajoz0	Barcelona0	Bilbao0	CiudadReal0	LasPalmas0	Madrid0	Madrid5	Madrid7	Murcia0	Oviedo0	Palma0	Pamplona0	Rioja0	Santander0	Santiago0	Sevilla0	Tenerife0	Valencia0	Valladolid0	Zaragoza0
2008-01	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-02	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-03	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-04	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-05	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-06	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-07	X	X	X	X	X	X	-	-	X	X	X	X	-	X	X	X	X	X	X	X
2008-08	X	X	X	X	X	X	-	-	X	X	X	X	X	X	X	X	X	X	X	X
2008-09	X	X	X	X	X	X	-	-	X	X	X	X	X	X	X	X	X	X	X	X
2008-10	X	X	X	X	X	X	-	-	X	X	X	X	X	X	X	X	X	X	X	X
2008-11	X	X	X	X	X	X	-	-	X	X	X	X	X	X	X	X	X	X	X	X
2008-12	X	X	X	X	X	X	-	-	X	X	X	X	X	X	X	X	X	X	X	X
2009-01	X	X	X	X	X	-	-	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-02	-	X	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-03	-	X	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-04	-	X	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-05	X	X	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-06	X	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-07	X	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-08	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-09	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-10	X	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-11	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2009-12	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-01	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-02	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-03	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-04	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-05	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-06	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-07	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-08	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-09	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-10	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-11	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2010-12	-	-	X	X	X	-	X	-	X	X	X	X	-	X	X	X	-	X	X	X

2011-01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2011-02	X	X	X	X	-	X	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2011-03	X	X	X	X	-	X	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2011-04	X	X	X	X	-	X	X	-	X	X	X	X	-	X	X	X	-	X	X	X
2011-05	X	X	X	X	-	X	X	-	X	X	X	X	-	X	X	/	-	X	X	X
2011-06	X	/	X	X	-	X	X	-	X	X	X	X	-	X	X	-	-	/	X	X
2011-07	/	-	/	/	-	/	/	-	/	/	/	/	-	/	/	-	-	-	/	/
2011-08	-	-	X	X	-	X	X	-	X	X	X	X	-	X	X	-	-	-	X	X
2011-09	X	-	X	X	-	-	X	-	X	X	X	X	-	X	X	-	-	-	X	X
2011-10	X	-	/	X	-	-	X	-	X	-	X	/	-	/	/	-	-	-	X	-
2011-11	X	-	-	-	-	-	X	-	-	-	X	-	-	/	-	-	-	-	X	-
2011-12	X	-	-	-	-	-	X	-	-	-	X	-	-	-	-	-	-	-	X	-
2012-01	X	-	-	-	-	-	X	-	-	-	X	-	-	-	-	-	-	-	X	-
2012-02	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	-	-	-	/	-
2012-03	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	-	-	-	X	-
2012-04	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	-	-	-	X	-
2012-05	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	-	-	-	X	-
2012-06	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	-	-	-	X	-
2012-07	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	-	-	-	X	-
2012-08	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	/	-	/	X	-
2012-09	-	-	-	-	-	-	-	-	-	-	X	-	-	-	-	X	-	X	X	-
2012-10	-	-	-	-	-	-	/	/	-	-	X	-	-	-	-	X	-	X	X	-
2012-11	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	X	-	X	X	-
2012-12	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	X	-	X	X	-
2013-01	-	-	-	-	-	-	X	-	-	-	/	-	-	-	-	/	-	/	/	/
2013-02	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2013-03	/	/	/	/	-	/	X	/	/	/	X	/	-	/	/	X	-	X	X	X
2013-04	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	X	-	X	X	X
2013-05	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	X	-	X	X	X
2013-06	-	/	/	/	-	/	X	/	/	/	X	/	-	/	/	X	-	X	X	X
2013-07	-	-	-	-	-	/	X	/	-	-	X	-	-	-	-	X	-	X	X	X
2013-08	/	/	/	/	-	/	X	/	/	/	X	/	-	/	/	X	-	X	X	X
2013-09	/	-	-	-	-	-	/	/	/	/	X	/	-	/	/	X	-	X	X	X
2013-10	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	X	-	X	X	X
2013-11	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	X	-	X	X	X
2013-12	-	-	-	-	-	-	X	-	-	-	X	-	-	-	-	X	-	X	X	X

Tabla A-1: Perspectiva de los archivos disponibles.

Nota: las equis (X) en la tabla corresponden a meses completos -considerando como completos aquellos con al menos veinte días de datos-; las barras (/), con meses con entre cinco y veinte días de datos aproximadamente; y los guiones (-) , con meses con menos de cinco días.

B FDAs del momento valle en cada PoP

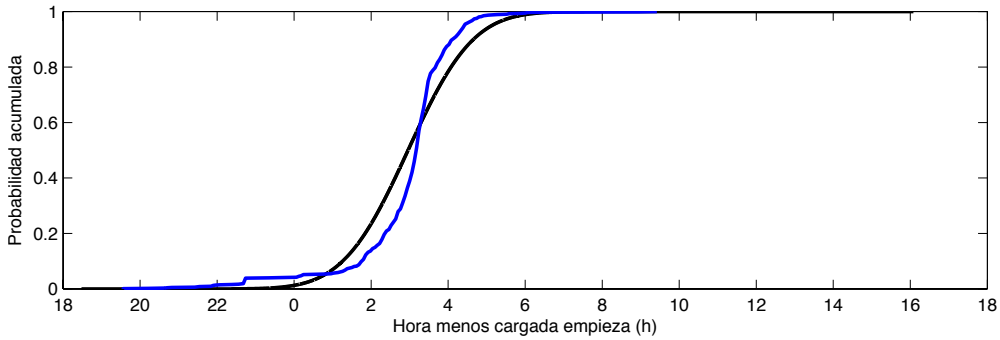

Figura B-1: FDA con desplazamiento correspondiente a PoP1

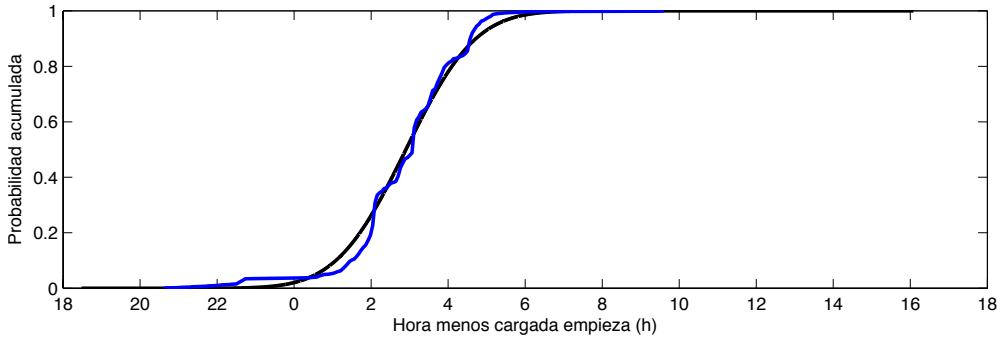

Figura B-2: FDA con desplazamiento correspondiente a PoP2

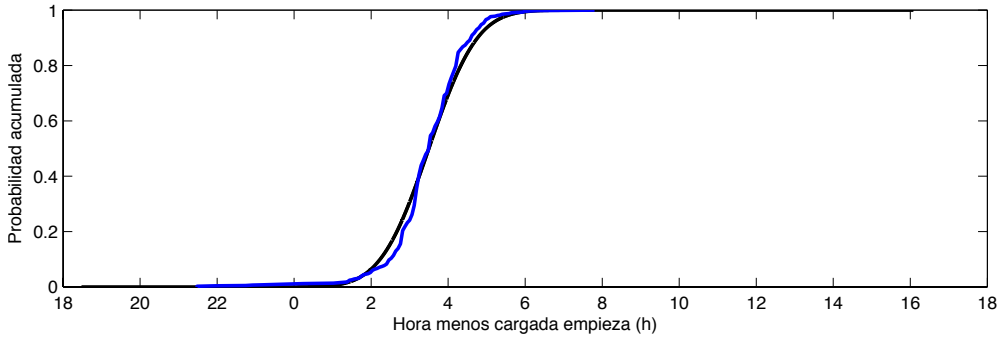

Figura B-3: FDA con desplazamiento correspondiente a PoP3

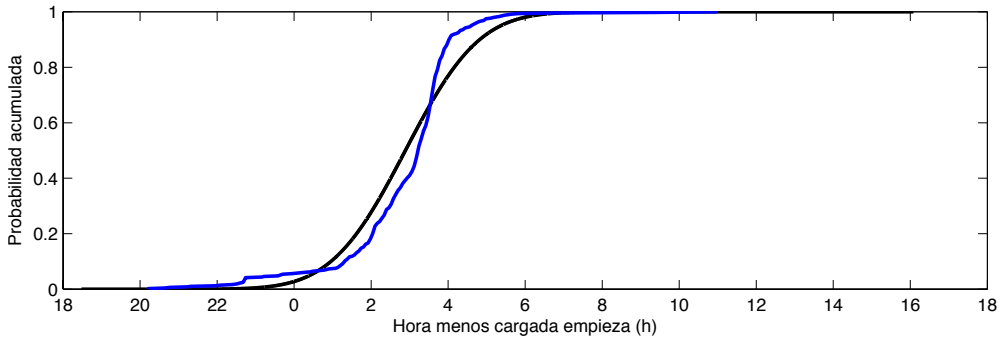

Figura B-4: FDA con desplazamiento correspondiente a PoP4

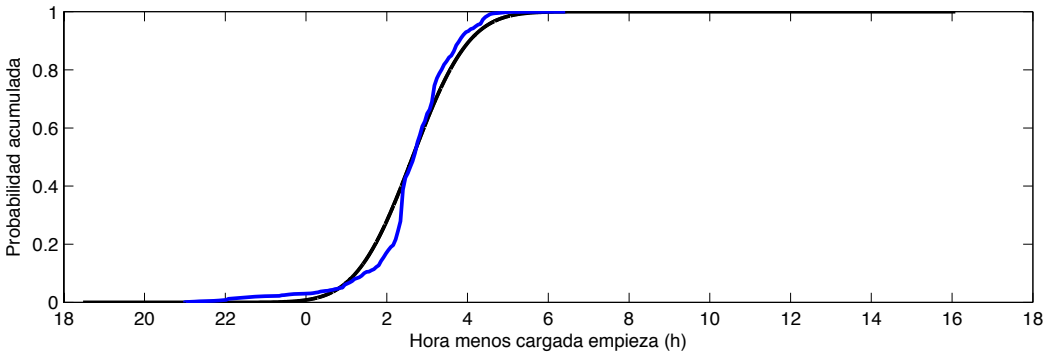

Figura B-5: FDA con desplazamiento correspondiente a PoP5

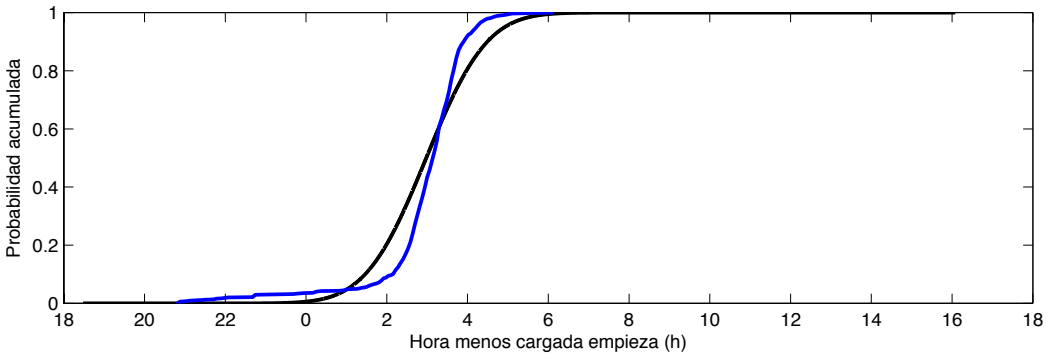

Figura B-6: FDA con desplazamiento correspondiente a PoP6

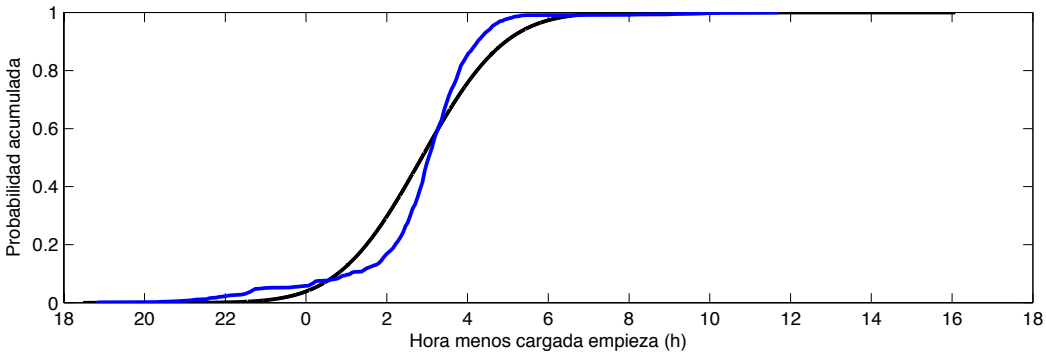

Figura B-7: FDA con desplazamiento correspondiente a PoP7

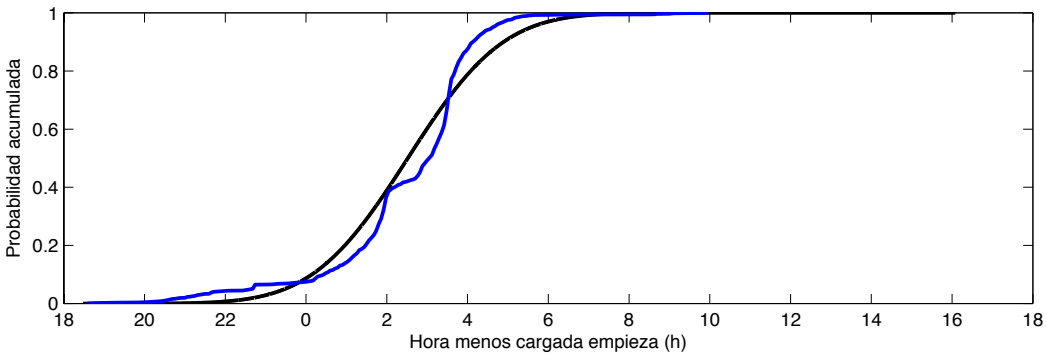

Figura B-8: FDA con desplazamiento correspondiente a PoP8

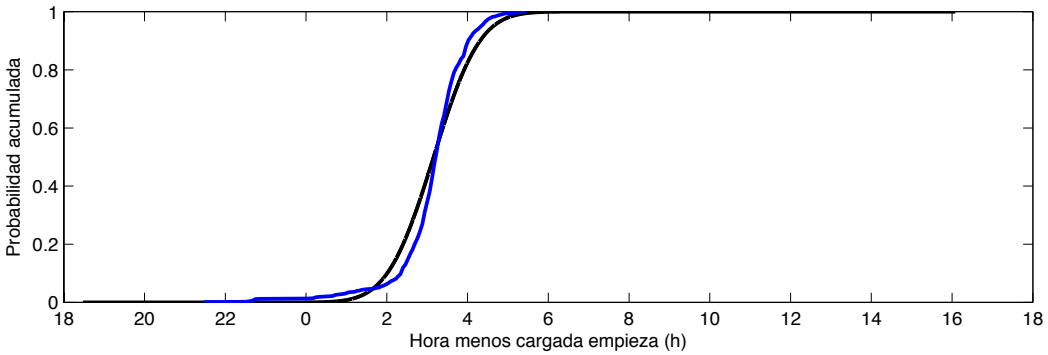

Figura B-9: FDA con desplazamiento correspondiente a PoP9

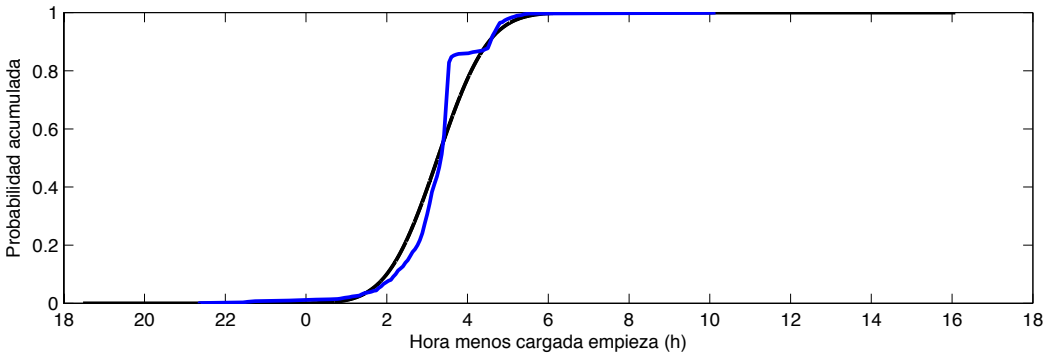

Figura B-10: FDA con desplazamiento correspondiente a PoP10

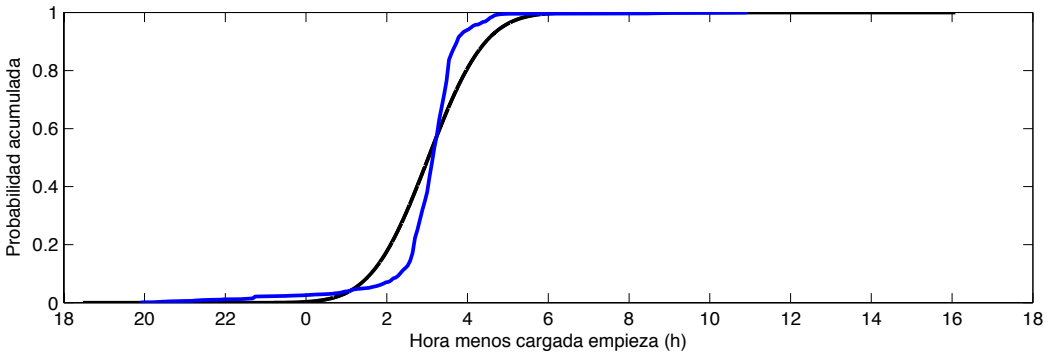

Figura B-11: FDA con desplazamiento correspondiente a PoP11

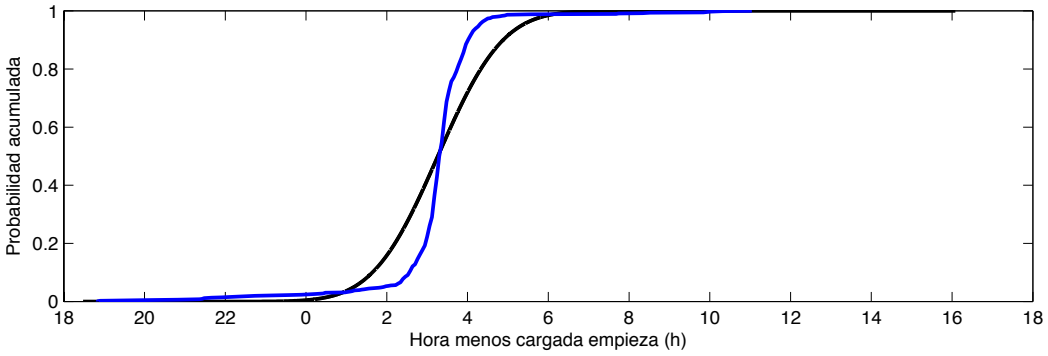

Figura B-12: FDA con desplazamiento correspondiente a PoP12

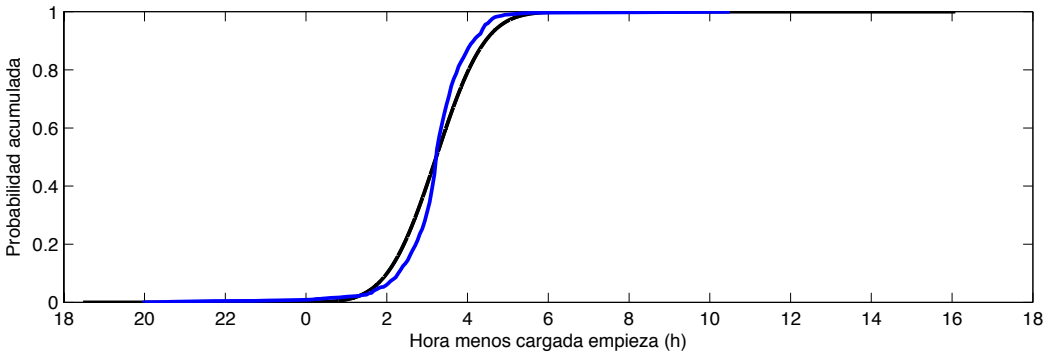

Figura B-13: FDA con desplazamiento correspondiente a PoP13

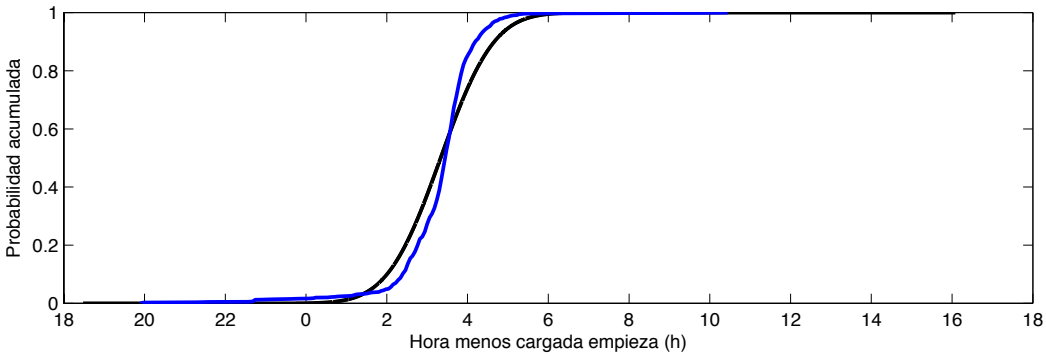

Figura B-14: FDA con desplazamiento correspondiente a PoP14